U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-256 310

Source Assessment: Severity of Stationary Air Pollution Sources A Simulation Approach

Monsanto Research Corp.

Prepared For
Industrial Environmental Research Lab.

July 1976

BEST SELLERS NTIS

FROM NATIONAL TECHNICAL INFORMATION SERVICE

Solar Heating and Cooling in Buildings: Methods of Economic Evaluation
COM-75-11070/PAT 48 p PC$4.00/MF$2.25

Design and Construction of a Residential Solar Heating and Cooling System
PB-237 042/PAT 233 p PC$8.00/MF$2.25

Performance of a Residential Solar Heating and Cooling System
PB-244 644/PAT 62 p PC$4.50/MF$2.25

Solar Energy
AD-778 846/PAT 478 p PC$12.50/MF$2.25

Stimulation of Geothermal Energy Resources
ERDA-37/PAT 52 p PC$4.50/MF$2.25

An Assessment of Industrial Energy Options Based on Coal and Nuclear Systems
ORNL-4995/PAT 329 p PC$10.00/MF$2.25

Environmental Impact Statements: A Handbook for Writers and Reviewers
PB-226 276/PAT 207 p PC$7.75/MF$2.25

Environmental Protection Guidelines for Construction Contract Specification Writers
ADA-014146/PAT 124 p PC$5.50/MF$2.25

A Review and Analysis of Environmental Impact Assessment Methodologies
ADA-013 359/PAT 21 p PC$3.50/MF$2.25

Procedures for Reviewing Environmental Impact Assessments and Statements for Construction Projects
ADA-015 020/PAT 25 p PC$3.50/MF$2.25

Computers in the 1980's: Trends in Hardware Technology
AD-783 323/PAT 20 p PC$3.50/MF$2.25

Minicomputers: A Review of Current Technology, Systems, and Applications
AD-783 316/PAT 104 p PC$5.50/MF$2.25

Microprocessor/Microcomputer Software Systems: Present and Future
ADA-013 322/PAT 25 p PC$3.50/MF$2.25

An Air Force Guide to Contracting for Software Acquisition
ADA-020 444/PAT 45 p PC$4.00/MF$2.25

A Dictionary for Unit Conversion
PB-249 659/PAT 456 p PC$17.50/MF Not Available

HOW TO ORDER

When you indicate the method of payment, please note if a purchase order is not accompanied by payment, you will be billed an addition $5.00 *ship and bill* charge. And please include the card expiration date when using American Express.

Normal delivery time takes three to five weeks. It is vital that you order by number or your order will be manually filled, insuring a delay. You can opt for *airmail delivery* for a $2.00 charge per item. Just check the *Airmail Service* box. If you're really pressed for time, call the NTIS Rush Order Service. (703) 557-4700. For a $10.00 charge per item, your order will be airmailed within 48 hours. Or, you can pick up your order in the Washington Information Center & Bookstore or at our Springfield Operations Center within 24 hours for a $6.00 per item charge.

You may also place your order by telephone or TELEX. The order desk number is (703) 557-4650 and the TELEX number is 89-9405.

Whenever a foreign sales price is NOT specified in the listings, all foreign buyers must add the following charges to each order: $2.50 for each paper copy; $1.50 for each microfiche; and $10.00 for each Published Search.

Thank you for your interest in NTIS. We appreciate your order.

METHOD OF PAYMENT

☐ Charge my NTIS deposit account no. ____________

☐ Purchase order no. ____________

☐ Check enclosed for $ ____________

☐ Charge to my American Express Card account number

☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Card expiration date ____________

Signature ____________

☐ Airmail Services requested

NAME ____________

ADDRESS ____________

CITY, STATE, ZIP ____________

Item Number	Quantity		Unit Price*	Total Price*
	Paper Copy (PC)	Microfiche (MF)		

*All Prices Subject to Change
10/76

Sub Total	
Additional Charge	
Enter Grand Total	

Clip and mail to:

NTIS
National Technical Information Service
U.S. DEPARTMENT OF COMMERCE
Springfield, Va. 22161
(703) 557-4650 TELEX 89-9405

237143

PB 256 310

EPA-600/2-76-032e
July 1976

Environmental Protection Technology Series

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL PROTECTION TECHNOLOGY series. This series describes research performed to develop and demonstrate instrumentation, equipment, and methodology to repair or prevent environmental degradation from point and non-point sources of pollution. This work provides the new or improved technology required for the control and treatment of pollution sources to meet environmental quality standards.

EPA REVIEW NOTICE

This report has been reviewed by the U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policy of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/2-76-032e	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE Source Assessment: Severity of Stationary Air Pollution Sources--A Simulation Approach		5. REPORT DATE July 1976 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) E. C. Eimutis, B. J. Holmes, and L. B. Mote		8. PERFORMING ORGANIZATION REPORT NO. MRC-DA-543
9. PERFORMING ORGANIZATION NAME AND ADDRESS Monsanto Research Corporation 1515 Nicholas Road Dayton, Ohio 45407		10. PROGRAM ELEMENT NO. 1AB015; ROAP 21AXM-071 11. CONTRACT/GRANT NO. 68-02-1874
12. SPONSORING AGENCY NAME AND ADDRESS EPA, Office of Research and Development Industrial Environmental Research Laboratory Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED Task Final; 6/75-5/76 14. SPONSORING AGENCY CODE EPA-ORD

15. SUPPLEMENTARY NOTES Project officer for this report is D.A. Denny, mail drop 62, 919/549-8411, ext 2547.

16. ABSTRACT The report gives results of a study simulating the establishment of the severity of stationary air pollution sources. The potential environmental impact of an emission source can be determined from the source severity (the ground level concentration contribution of pollutants relative to some potentially hazardous concentration of the same species). The frequency distribution of the severity of well-documented source types can be examined deterministically. A statistical approach is required to simulate the frequency distribution of the severity of source types that are complex or involve a large number of emission points in order to ultimately assess such sources. A Monte Carlo simulation technique is described in this report, together with efficient algorithms for fitting the inverse Weibull, gamma, normal, and log-normal cumulative density functions. Significant correlation is demonstrated between deterministic and simulated severity results using coal-fired steam/electric utilities as an example.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS		b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution	Electric Utilities	Air Pollution Control	13B
Ranking		Stationary Sources	12B
Environmental Biology		Source Assessment	06F
Simulation		Environmental Impact	
Monte Carlo Method			12A
Estimating			

18. DISTRIBUTION STATEMENT Unlimited	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 133
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE 6.00

EPA Form 2220-1 (9-73)

PRICES SUBJECT TO CHANGE

EPA-600/2-76-032e
July 1976

SOURCE ASSESSMENT:

SEVERITY OF STATIONARY AIR POLLUTION SOURCES--A SIMULATION APPROACH

by

E.C. Eimutis, B.J. Holmes, and L.B. Mote

Monsanto Research Corporation
1515 Nicholas Road
Dayton, Ohio 45407

Contract No. 68-02-1874
ROAP No. 21AXM-071
Program Element No. 1AB015

EPA Project Officer: Dale A. Denny

Industrial Environmental Research Laboratory
Office of Energy, Minerals, and Industry
Research Triangle Park, NC 27711

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Research and Development
Washington, DC 20460

PREFACE

The Industrial Environmental Research Laboratory (IERL) of EPA has the responsibility for insuring that air pollution control technology is available for stationary sources. If control technology is unavailable, inadequate, uneconomical or socially unacceptable, then development of the needed control techniques is conducted by IERL. Approaches considered include process modifications, feedstock modifications, add-on control devices, and complete process substitution. The scale of control technology programs ranges from bench to full scale demonstration plants.

The Chemical Processes Branch of IERL has the responsibility for developing control technology for a large number (>500) of operations in the chemical and related industries. As in any technical program the first step is to identify the unsolved problems.

Each of the industries is to be examined in detail to determine if there is sufficient potential environmental risk to justify the development of control technology by IERL. Monsanto Research Corporation (MRC) has contracted with EPA to investigate the environmental impact of various industries which represent sources of emissions in accordance with EPA's responsibility as outlined above. Dr. Robert C. Binning serves as Program Manager in this overall program entitled, "Source Assessment." As a first step, MRC has developed a priority listing of the industries in each of four categories: combustion, organic materials, inorganic materials, and open sources. The purpose and intended use of this listing is that it serve as one of several guides to the selection of those sources for which MRC will perform detailed source assessments. Source assessment documents will be produced by MRC and used by EPA to make decisions regarding the need for developing additional control technology for each specific source.

In order to analyze the severity of those sources in which the emission points number in the thousands or hundred thousands, a statistical approach is required such as the Monte Carlo simulation technique described in this report. An example of this approach for analyzing coal-fired steam electric utilities is included.

CONTENTS

Section		Page
I	Introduction	1
II	Source Severity	3
	A. Mathematical Structure	5
	B. Derivation of $\overline{X}$	6
	C. Pollutant Severity Equations	7
III	Simulation Methodology	9
	A. Introduction	9
	B. Theory and Methodology	12
	1. All Input Variables Independent	13
	2. Dependent Input Variables	15
IV	Example of Use of Simulation Approach With Coal-Fired Electric Utilities	19
V	Discussion of Non-Normality and Chi-Square Goodness-Of-Fit Test	25
	A. Central Limit Theorem and T-Test	26
	B. Coefficient of Skewness and Kurtosis as Analytical Measures of Non-Normal Distributions	34
	C. Weibull Distribution	37
	D. Gamma Distribution	41
	E. Log-Normal Distribution	44
	F. Sample Skewness and Kurtosis	45
	G. The Chi-Square Goodness-Of-Fit Test	49
VI	Appendixes	57
	A. Standard Statistical Formulae For Finite Populations	58
	B. Detailed Derivations of the Criteria Pollutant Severity Equations	61
	C. The Simulation Programs	66
	D. The Goodness-Of-Fit Program	94
	E. Treatment of Correlated Data by Linear Transformation	111
VII	References	117

Preceding page blank

LIST OF TABLES

Table		Page
1	Criteria Pollutant Severity Equations	8
2	Results of Deterministic Calculations	20
3	Notation Used for Statistical Parameters	27
4	Comparison of Random Sample Values and Population Mean for Coal-Fired Electric Utilities	33
5	Values for Various Points of Interest in the Weibull Distribution for Various Values of Parameter b (a=1)	40
6	Values for Various Points of Interest in the Weibull Distribution for Various Values of Parameter b ($a=1.0 \times 10^{-5}$)	41
7	0.05 and 0.01 Points of the Distribution of γ_1, the Coefficient of Skewness	47
8	Percentage Points of the Distribution of γ_2, the Measure of Kurtosis	48
9	Values of G_1, G_2 and (G_2-3) for Power Plant Example	49
10	Theoretical and Actual Frequencies for Nine Class Frequencies	53
11	Theoretical and Actual Frequencies for Nine Class Intervals	54
A-1	First Square Root Factor in Equation A-3 as a Function of Sample Size	59
C-1	Variables, Distributions, Parameters and Clips for Coal-Fired Electric Utilities Example	78
C-2	Summary of Input Data by Groups for Coal-Fired Electric Utilities Example	79
C-3	Computer Listings of the Simulation Programs	82
D-1	Theoretical and Actual Frequencies for Various Class Intervals	99
D-2	Computer Listings of the Simulation Programs Programs - Goodness-Of-Fit Program	101
E-1	Comparison of Simulation Results	113

LIST OF FIGURES

Figure		Page
1	Frequency Histograms for the Severity of SO_2 Emissions from Coal-Fired Electric Utilities Comparing the Simulation and Deterministic Methods	21
2	Comparison of Cumulative Frequency for the Severity of SO_2 Emissions from Coal-Fired Electric Utilities Using the Simulation and Deterministic Methods	22
3	Unimodal Continuous Distribution	35
4	Examples of Kurtosis in Distributions	36
5	Probability Density Function for the Exponential Distribution	37
6	Probability Density Function of the Weibull Distribution for Various Values of Parameter b	38
7	Probability Density Function for the Gamma Distribution for Various Values of Parameter α	43
8	Probability Density Function for the Log-Normal Distribution	45

LIST OF SYMBOLS

Symbol	Definition
A	Y-intercept for regression line [=YB - B(XB)]
A,B,C,D,E...	Simulated variable values, same as IVAL where A = variable 1, B = variable 2, etc
a, b, α, β	Arbitrary parameters
B	Slope of regression line $\left(= \frac{R \cdot SY}{SX}\right)$
B'	Average breathing rate
CC	Coal consumed
c.d.f.	Cumulative distribution function
C_1, C_2, CLIP(I,J)	Two dimensional array; maximum/minimum values for associated stochastic variables
E	Emission factor
e	2.72
F	Hazard potential factor (=TLV·K) for severity calculations, or statistical ratio of variances (Appendix E)
$F_1, \ldots, F_n$	Actual frequencies of the sample data in n class intervals
$f_1, \ldots, f_n$	Theoretical frequencies that would be expected for a sample of the same size from the given distribution
g_1	Coefficient of skewness of sample
g_2	Measure of kurtosis of sample
H	Relative hazard
h	Emission height
H_0	Hypothesis that data are distributed according to some given distribution
ICØDE	Two dimensional array containing information on which dependent variable is correlated and which distribution it fits
ICØDE(1)	Code which identifies the type of distribution for independent variable 1
ITIL	Title for x-axis on plot and title of report output
IVAL	Variable value, selected from VAR(I,J)

Preceding page blank

LIST OF SYMBOLS (Continued)

Symbol	Definition
K	Safety factor = $\left(\frac{8}{24}\right)\left(\frac{1}{100}\right)$
$K_1, K_2 \ldots$	Conversion factors
LD_{50}	Lethal dose for 50% of the people exposed
LT	Input integer used to indicate specific option to be used in program
m_3, m_4, mk	Third, fourth, kth central moment about $\bar{X}$, respectively
N	Population exposed to a specific source, persons
n	Number of samples or class intervals (for simulation) or number of point sources emitting the ith material (otherwise)
NCFLAG	Input integer which provides the option of using Sturge's rule to set up the number of class intervals and using a value of XMIN and XMAX to establish W and the resulting class intervals
NDVAR	An integer which gives the number of dependent input variables
NFLAG	Flag to identify whether or not entered data are last data set to be analyzed
NGRØUP	An integer which tells the simulation program the sample size for each pass
NINT	The number of class intervals to be used in program
NIVAR	An integer which tell the simulation program how many input variables are independent
NPASS	An integer which tells the simulation program how many passes to make
NSAMP	Sample size
P	Probability
PAR(I,J)	Paramter matrix with values for each statistical distribution
p.d.f.	Probability density function
PNUM(I)	Same as SNUM(1) but for population prediction

LIST OF SYMBOLS (Continued)

Symbol	Definition
Q	Emission rate
R	Linear correlation coefficient or random number (Appendix C)
R'	Lung retention factor for the pollutant of interest (dimensionless factor, $0<R'<1$)
S	Source severity or standard deviation corrected for small sample size
SD	Standard deviation of Y for a given X
$(SD)^2$	Sample variance
SE	Standard error $(= SY\sqrt{1 - R^2})$
SF	Subprogram to calculate severity
S_{ij}	Source severity of the $i^{\underline{th}}$ material in the region around the $j^{\underline{th}}$ source
SNUM(I)	Resultant values of severity, where I=1 indicates the probability that S lies in the range <0.1, I=2 indicates $0.1<S<1$, and I=3 indicates $S>1$
$\hat{S}$,SD	Sample standard deviation
SX	Standard deviation of X
$S_{\overline{X}}$	Standard deviation of $\overline{X}$
SY	Standard deviation of Y
S_1, S_2	Sample standard deviation of X1 and X2 data, respectively
T	The "Student t" random variable with n degrees of freedom $(=Z/\sqrt{\chi^2/n})$
TLV	Threshold limit value
T'	Total time $(=t_2-t_1)$
t	Time
t_1	Start time
t_2	Finish time
u	Wind speed
VAR_1, VAR_2, VAR_3	Independent variables
W	Width of class intervals for a histogram
X	Stochastic variable
$\overline{X}$	Sample mean $\left(= \frac{\Sigma x}{n}\right)$

LIST OF SYMBOLS (Continued)

Symbol	Definition
$\overline{X}_n$	Mean of a sample of size n
XB	Mean deviation of X
XMAX	Maximum value
XMIN	Minimum value
XPØP	Population size
XSAMP	Actual sample size
$X_1,\ldots X_n$	A random sample from the population
X1,X2	Arrays of correlated data
YB	Mean deviation of Y
Y1,Y2	Arrays of transformed data
Z	A standard normal random variable
z	Random variable $[z = f(x_1,\ldots,x_n)]$
α	Angle of rotation in radians to make correlation zero
Γ	Gamma function
γ_1	Coefficient of skewness of population
γ_2	Measure of kurtosis of population
ϵ	Any positive member of population
μ	Population mean
$\hat{\mu}$	Estimate of the total population mean
μ_3,μ_4	Third and fourth central moment about μ, respectively
$\xi 1$	Mean of Y1 (=XI1)
$\xi 2$	Mean of Y2 (=XI2)
π	3.14159
ρ	Correlation coefficient
σ	Standard deviation
$\hat{\sigma}$	Estimate of population standard deviation
σ^2	Population variance
$\hat{\sigma}_{\overline{x}}$	Estimate of standard error of mean
σ_y	Horizontal dispersion coefficient
σ_z	Vertical dispersion coefficient
$\overline{\chi}$	Average maximum ground level concentration

LIST OF SYMBOLS (Continued)

Symbol	Definition
χ^2	Any Chi-square random variable with n degrees of freedom
χ_{max}	Maximum ground level concentration
$\chi(t)$	Concentration time history
Ψ	Delivered dose = $B'\cdot R'\cdot\int\chi(t)dt$
Ψ_A	Actual pollutant dose delivered from a given point source $(=N\cdot B'\cdot R'\cdot T'\cdot\chi)$
Ψ_F	Potentially hazardous dose $(=N\cdot B'\cdot R'\cdot T'\cdot F)$

SECTION I

INTRODUCTION

A prioritization listing of air pollution sources was developed earlier to serve as a first step in selecting industries for detailed study to determine whether a sufficient potential environmental risk exists to justify the development of control technology. Preparation of the listing or relative ranking of emission sources involved the use of an impact factor which is a worst case, integral quantity. In current assessment studies, one of the potential environmental impacts of a source is determined from the source severity, defined as the ground level concentration contribution of pollutants relative to some potentially hazardous concentration of the same species. The frequency distribution of the severity of well-documented source types can be examined deterministically. However, source types which are complex or involve a large number of emission points require a statistical approach to simulate the frequency distribution of the severity and ultimately permit the assessment of the source.

The basic premise of the simulation approach is that detailed information on all required factors and all emission points for certain source types will not be obtained because of time or cost limitations. When such detailed information cannot be collected, the inputs can be described in terms of a distribution of values. In many cases (e.g., electric utility boiler capacities), these distributions are readily available; for some sources, special approximating techniques must be used to develop them. Having developed frequency distributions for the inputs, a distribution of severity can be calculated by means of a simulation technique. When several input variables are treated as frequency distributions, the computation is extremely tedious; however, with a high-speed digital computer, computation times are on the order of a few seconds.

This document presents a methodology for describing the severity distributions of various source types. A Monte Carlo simulation technique is described together with efficient algorithms for fitting the inverse Weibull, gamma, normal, and log-normal cumulative density functions. Using coal-fired steam electric utilities as an example, significant correlation is demonstrated between deterministic and simulated severity results.

SECTION II

SOURCE SEVERITY

The air pollution severity, S, of a given source should in some way be proportional to the degree of potential hazard it imposes upon individuals in its environment. The relative hazard, H, from a specific emission can be defined as being directly proportional to the delivered dose, the probability of dose delivery, and number of people who would receive it, and inversely proportional to the toxicity of the material as follows:

$$S \propto H \propto \frac{NP\Psi}{LD_{50}} \qquad (1)$$

where S = source severity
H = relative hazard
N = number of persons
LD_{50} = lethal dose for 50% of the people exposed
P = probability of dose delivery
Ψ = delivered dose = $B' \cdot R' \cdot \int \chi(t)dt$
B' = average breathing rate
R' = lung retention factor
$\chi(t)$ = concentration time history

The source severity is herein defined as the ratio of the dose of a pollutant delivered to a population, relative to some potentially hazardous dose. Since LD_{50} data are not available for human beings, another measure of potentially hazardous dosage was used.

The potentially hazardous dose for a given pollutant from a specific point source in a given region is defined as follows:

$$\Psi_F = NB'R' \int_{t_1}^{t_2} TLV(t)\ K\ dt \qquad (2)$$

where Ψ_F = potentially hazardous dose, g

N = population exposed to a specific source, persons

B' = average breathing rate, m^3/s-person

R' = lung retention factor for the pollutant of interest (dimensionless factor, $0<R'<1$)

K = safety factor = $\left(\frac{8}{24}\right)\left(\frac{1}{100}\right)$

t = time

t_1 = start time, s

t_2 = finish time, s

TLV® = threshold limit value, g/m^3

The total time of interest, T', is defined as:

$$T' = t_2 - t_1 \qquad (3)$$

Similarly, a hazard potential factor, F, is defined as:

$$F = TLV \cdot K \qquad (4)$$

Since TLV is a constant,

$$\Psi_F = N \cdot B' \cdot R' \cdot T' \qquad (5)$$

The actual pollutant dose delivered, Ψ_A, from a given point source can be calculated as follows:

$$\Psi_A = N \cdot B' \cdot R' \int_{t_1}^{t_2} \chi(t)\, dt \qquad (6)$$

where $\chi(t)$ = the actual ground level concentration time history of a pollutant of interest emitted by a specific point source, g/m^3

The value of $\chi(t)$ is very difficult to obtain and was therefore approximated by an average value, $\bar{\chi}$. The total actual dose delivered for a specific pollutant from a specific source is then:

$$\Psi_A = N \cdot B' \cdot R' \cdot T' \cdot \bar{\chi} \qquad (7)$$

Since our measure of source severity was defined as the ratio of the two dosages, then:

$$S = \frac{\Psi_A}{\Psi_F} = \frac{N \cdot B' \cdot R' \cdot T' \cdot \bar{\chi}}{N \cdot B' \cdot R' \cdot T' \cdot F} \qquad (8)$$

or

$$S = \frac{\bar{\chi}}{F} \qquad (9)$$

A. MATHEMATICAL STRUCTURE

The source severity, S, of the $i^{\underline{th}}$ material in the region around the $j^{\underline{th}}$ source is expressed as the ratio of $\bar{\chi}_{ij}$ to F_i; i.e.,

$$S_{ij} = \frac{\bar{\chi}_{ij}}{F_i} \qquad (10)$$

For the $i^{\underline{th}}$ emitted material, the severity vector, S_i, is defined by:

$$S_i = \begin{pmatrix} S_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ S_{in} \end{pmatrix} \tag{11}$$

where n = number of point sources emitting the $i^{\underline{th}}$ material

The mean and median severity for the $i^{\underline{th}}$ material may be obtained from a cumulative frequency plot of S_i. In addition, n-fractiles, the standard deviation, and confidence limits about the mean severity can be calculated. Since all of the populations under consideration are finite, alternate forms of the standard statistical equations were used (as presented in Appendix A).

B. DERIVATION OF $\bar{\chi}$

Since source to receptor distances were not compiled, the maximum ground level concentration for elevated point sources, χ_{max},[1] was used:

$$\chi_{max} = \frac{2\ Q\sigma_z}{\pi e u h^2 \sigma_y} \tag{12}$$

where χ_{max} = maximum ground level concentration, g/m^3
π = 3.14159
e = 2.72
u = wind speed, m/s

---

[1] Slade, D. H. (ed.). Meteorology and Atomic Energy. Environmental Science Services Administration, Air Resources Labs. Silver Spring. AEC Publication No. TID-24190. July 1968. 445 p.

h = emission height, m
σ_z = vertical dispersion coefficient, m
σ_y = horizontal dispersion coefficent, m
Q = emission rate, g/s

The average concentration, $\bar{\chi}$, is a function of sampling time, t, and it can be related to the maximum concentration χ_{max} as follows:[2]

$$\bar{\chi} = \chi_{max} \frac{t_1}{t_2}^{p} \qquad (13)$$

where t_1 = 3 min
t_2 = reference time period, min
p = 0.17

C. POLLUTANT SEVERITY EQUATIONS

For any material with a given TLV, the severity equation becomes:

$$S = \frac{\bar{\chi}}{F} = \frac{\chi_{max}\left(\frac{t_1}{t_2}\right)^{p}}{TLV\left(\frac{8}{24}\right)\left(\frac{1}{100}\right)} \qquad (14)$$

Assuming a national average wind speed of approximately 10 mph (4.5 m/s), an averaging period of 24 hours and substituting the appropriate values, Equation 14 becomes:

$$S = \frac{\bar{\chi}}{F} = \frac{\chi_{max}\left(\frac{3}{1440}\right)^{0.17}}{(TLV)(3.33 \times 10^{-3})}$$

$$= \frac{0.35\ \chi_{max}}{(TLV)3.33 \times 10^{-3}}$$

$$= \frac{105\ \chi_{max}}{TLV}$$

---

[2] Turner, D. B. Workbook of Atmospheric Dispersion Estimates, 1970 Revision. U.S. Department of Health, Education, and Welfare. Cincinnati. Public Health Service Publication No. 999-AP-26. May 1970. 84 p.

or

$$S = \frac{(2)(105)Q\sigma_z}{\pi e u h^2 \sigma_y (TLV)} \qquad (15)$$

The national average atmospheric stability is approximately neutral. Hence, $\sigma_y \cong \sigma_z$, and:

$$\frac{\sigma_z}{\sigma_y} \cong 1.0 \qquad (16)$$

Thus:

$$S = \frac{210Q}{38.43h^2(TLV)}$$

or,

$$S = \frac{5.5Q}{(TLV)h^2} \qquad (17)$$

Since the criteria pollutants (particulates, SO_x, NO_x, CO, and hydrocarbons) have established ambient air standards, the appropriate standard (in g/m^3) can be substituted for the potential hazard factor, F. The severity equations for the five criteria pollutants are shown in Table 1. (Detailed derivations are shown in Appendix B).

Table 1. CRITERIA POLLUTANT SEVERITY EQUATIONS

Pollutant	Severity equation	
Particulate	$S = 70Qh^{-2}$	(18)
SO_x	$S = 50Qh^{-2}$	(19)
NO_x	$S = 315Qh^{-2.1}$	(20)
Hydrocarbons	$S = 162.5Qh^{-2}$	(21)
CO	$S = 0.78Qh^{-2}$	(22)

SECTION III

SIMULATION METHODOLOGY

A. INTRODUCTION

In many statistical analyses of data, it is frequently desired to consider a random variable which is a function of other random variables. An example pertinent to air pollution studies is given by the severity equations for ground level concentrations of air pollutants. For example, the severity equation for SO_2 emissions from the stacks of coal-fired electric utility plants is given by:

$$S = \frac{50Q}{h^2} \qquad (19)$$

where Q = emission rate, g/s
h = emission height, m

The emission rate can be calculated from:

$$Q = (CC)(E)(\%\ \text{sulfur})(K_1) \qquad (23)$$

where CC = coal consumed, g/yr

$$E = \text{emission factor} = \frac{0.019\ \text{g}\ SO_2\ (1\%\ \text{sulfur coal})}{\text{g of coal consumed}}$$

% sulfur = percent of sulfur in the coal

$K_1 = 3.171 \times 10^{-8}$ (to convert from g/yr to g/s)

or

$$S = \frac{(K_2)(CC)(\%\ \text{sulfur})}{h^2} \qquad (24)$$

where $K_2 = 3 \times 10^{-9}$

Next, consider a general setting where the random variable z is a function of the random variables $x_1, \ldots, x_n$ given by $z = f(x_1, \ldots, x_n)$ for some function f. Suppose the actual distributions of the input random variables $x_1, \ldots, x_n$ are known including their probability density functions (p.d.f.) and the corresponding cumulative distribution functions (c.d.f.). Then it seems reasonable to assume that the distribution of the random variable z can be obtained. In a sense this is true in that integral formulae have been developed which give the probability density function and the cumulative distribution function for z as a function of the same functions for the x_i.[3] These formulae, however, are very complicated even for the case of the simple sum, difference, product, or quotient of two random variables. Also, even if the integrals are successfully evaluated, the resulting probability density function for z will in general not be exactly one of the standard distributions and as a result may be difficult to handle. There are certain special cases in which the resulting p.d.f. will be known.[3] In these instances, the analytical approach to finding z explicitly is by far the best approach. In other instances certain simplifying assumptions about the distribution of z can be made provided certain things are true about the coefficient of variability or equivalently the coefficient of skewness of the input variables.[4] However, in cases where there are more than two input variables or there is considerable skewness exhibited by the variables or the function f becomes complicated, then the strict analytical approach to finding the distribution of z explicitly will in general not be applicable.

---

[3] Parzen, E. Modern Probability Theory and Its Applications. New York, John Wiley & Sons, 1960.

[4] Springer, C. H., et al. Probabilistic Models. Homewood, Richard D. Irwin, Inc., 1968.

In these cases, where the general approach of finding the explicit distribution function for z is not applicable, an alternate approach is to calculate "many" values of z for explicit values of the input variables $x_1, \ldots, x_n$ and use these values to estimate (rather closely if enough values of z are known) such things as the mean, standard deviation, etc., for z. Also, class intervals can be formed and a frequency histogram and cumulative distribution plot can be developed for the "many" calculated values of z. This will yield a distribution of z without any knowledge of an analytical formula for its p.d.f. or even without knowing whether any of the known standard distributions of statistics exist for the distribution. This approach is called the deterministic approach because in this technique it is possible to determine explicit values for z from explicit values of the input variables $x_1, \ldots, x_n$. This approach is an approximation to the strictly analytical approach described earlier. The deterministic approach works well whenever it is possible to actually calculate the "many" values of z deterministically from given values of the input variables. The term "many" means at least 30 values for the purpose of estimating the mean, standard deviation, and a 95% confidence interval for the mean. (The t-test for finding confidence intervals is discussed in Section V of this report.) However, to obtain a better frequency histogram for z, 100 or more values of z should be available. (A histogram can be constructed with less values but it will tend to be less meaningful because the number of class intervals will have to be smaller.)

Finally consider the situation when either no explicit values of the input variable are available from which values of z can be calculated or the number of such values is too small to permit calculation of enough values of z to determine useful information regarding its distribution. In this situation a tool commonly used is the probabilistic approach which uses a computer simulation to obtain values for z. For example, instead of knowing many values for the input variables $x_1, \ldots, x_n$, only limited information may be available, such as an estimate of the mean and possibly the range and symmetry or skewness properties. Such situations are not amenable to either the analytical or deterministic approach. In this case, the input variables are fitted to some theoretical distribution and the small amount of available information about the variables is used to determine the parameters of the distributions. The computer is then used to sample a large number of times from each input variable's distribution function and to subsequently use these data to calculate a large number of values of z from which the mean, standard deviation, etc., can be estimated and frequency histograms and cumulative distribution plots for z can be prepared. Some of the techniques and procedures used in such a computer simulation are described below.

B. THEORY AND METHODOLOGY

The equation for the severity (S) of ground level concentrations of SO_2 emissions from the stacks of coal-fired electric utilities will be used to illustrate the methodology utilized in the simulation approach. The severity equation is:

$$S = \frac{(K_2)(CC)(\%\ \text{sulfur})}{h^2} \qquad (24)$$

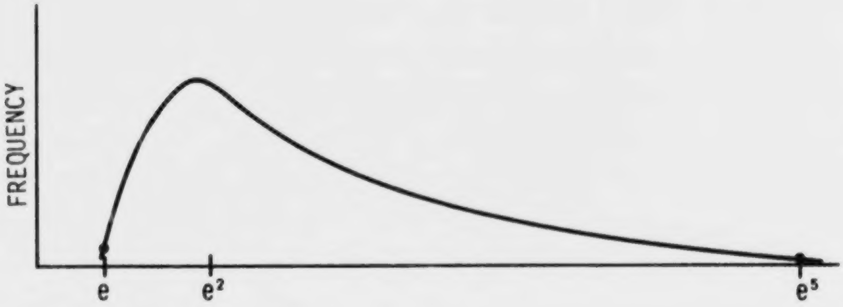
where K_2, % sulfur, CC, and h have the meanings defined earlier. Thus, the input random variables are % sulfur, CC, and h.

1. All Input Variables Independent

When all of the input random variables are independent, the methodology is relatively simple. A large sample (e.g., of size n) is drawn from the distribution of each of the input variables. These data are then used one by one to calculate n values of S. From these n values the mean, standard deviation, etc., can be calculated and a frequency histogram and cumulative distribution can be plotted.

Some comments are in order regarding the method by which samples are drawn from the distribution of the input variables. First, it should be noted that the input variables are restricted to one of four types of continuous distributions: the Weibull, normal, gamma, or log-normal distribution. The type of each input variable and the corresponding parameters for its distribution function must thus be specified. The method of obtaining the "best" type for each variable and the corresponding parameters is described in Section V and the parameters for the SO_2 example are given in Appendix C. In these goodness-of-fit procedures, it is necessary to have a random sample of data points for the input variable in order to be able to fit it to the proper distribution. However, certain situations may arise when that much information about the input variable is not available. For example, two extreme points on the distribution and either the mean or mode may be known, or some information may be available to determine whether the distribution is symmetric or skewed. In such situations where the goodness-of-fit program is inoperable, it may still be possible to fit the variable to one of the four distributions above and to obtain its parameters.

As a demonstration of the above procedure, consider the following example. Suppose that for an input variable, x, it is known with 95% confidence that the values of x will be between e and e^5 (where e = 2.7..). Suppose also that the mode of the distribution is known to be between e and e^2 and that the mean is approximately equal to e^3. These points then indicate that x is a rather heavily skewed right distribution. The graph for the p.d.f. of x may resemble the one shown below:



Since it is known that the 0.025 point on the cumulative graph is approximately equal to e and the 0.975 point is approximately equal to e^5, this information can be used alone to calculate A and B in a Weibull fit (described later). Thus, one finds that A = 1.25 and B = 7.29×10^{-3}. These values of A and B yield a theoretical mean $\mu = 47.7$ which is larger than the estimated e^3 value for the mean. The theoretical mode is 14.2 which again is larger than the estimated mode. Thus, the Weibull fit could be used as an approximation to the "true" distribution of x, although it is not a very good fit as our observations about the mean and mode indicate.

Another way of obtaining a distribution for x is to assume that it is log-normally distributed since the log-normal distribution is a right-skewed distribution. If x is assumed to be a log-normal distribution, then log x must be normal. Hence, by taking the logarithm of the 0.025 point and 0.975 point of x, the same points on the cumulative graph of log x are obtained which were assumed to be normal. These points are 1 and 5, respectively. Thus, the mean μ of log x should be taken to be 3 and, since 1 and 5 are the 0.025 and 0.975

points, respectively, it is found that $\sigma = 1.2$. The values $\mu = 3$ and $\sigma = 1.2$ can thus be used as parameters to sample from the normal for values of log x. By taking antilogarithms of the sample, a sample for x can be obtained.

In view of the above discussion, it is evident that several avenues are available for obtaining a distribution to fit the given data or information about each input variable. The simulation program (for the case of independent input variables) simply takes the parameters for the given type of distribution for an input variable and samples from this distribution to obtain a random sample for that input variable.

The method by which the program selects the sample from a given distribution varies with the distribution. The direct approach is used for the Weibull and normal distributions. The rejection method is used for the gamma distribution. Finally, for the log-normal distribution, the direct approach is used to sample from the normal with the mean equal to the mean of log x and the standard deviation equal to the standard deviation of log x. These are sometimes obtained by taking the log of the geometric mean and geometric standard deviation of x. After obtaining a sample for log x, the antilogarithm of these values gives a sample for x. These methods of sampling are further discussed later in this report.

2. Dependent Input Variables

Consider what happens when two (or more) of the input variables are correlated to some degree. If this situation occurs, and a sampling procedure is used such as the one discussed above, which assumes independent variables, it will tend to distort the mean as well as the distribution of the output variable. For example, in the severity example it was found that the variables CC and h had a sample correlation coefficient of about 0.55. Thus, whenever CC and h were sampled independently, large values of CC were allowed with small values of h and vice versa. This procedure tended to distort the "true" distribution of S by allowing unrealistically low values for S and, more importantly, unrealistically high values. These high values caused the simulated mean to be 16.0 whereas the true mean was 8.9 for the deterministic population calculation. Thus, some way was needed to account for the correlation that existed between CC and h.

Consider two input variables, X and Y, which are correlated with linear correlation coefficient R. This value of R can either be estimated and supplied directly to the program or it can be obtained by a simple regression on the sample data for X and Y. Once R is obtained, the slope, B, and the Y-intercept, A, for the regression line is given by:

$$B = \frac{R \cdot SY}{SX} \tag{25}$$

$$A = YB - B(XB) \tag{26}$$

where XB and SX are the mean and standard deviation of X, and YB and SY are the corresponding parameters for Y. (This assumes that X is taken as the independent variable in the regression line.) If R had been supplied directly to the program (without any sample data), then XB, SX, YB, and SY would also be required in order to calculate A and B.

After A and B are obtained, the following relationship exists:

$$Y = A + BX = YB + B(X - XB) \tag{27}$$

with an error term to account for the fact that the regression line is not an exact relation between X and Y. The next item needed is the standard deviation (or error), SE, of Y due to the regression line estimate of Y. This is given by:

$$SE = SY\sqrt{1 - R^2} \tag{28}$$

Whenever R = 0 (indicating that X and Y are independent), then SE = SY as would be expected (i.e., the standard deviation of Y due to the regression line estimate is simply SY). Also, if R = ±1, then SE = 0 as would be expected since there is no deviation from the regression line in this case.

SE shown above is the standard deviation for Y when the regression line estimate of Y is used and the value of X is XB, which is not likely to occur often. Hence, a way to compute the standard deviation (SD) of Y is needed using the regression estimate of Y and any value of X. Intuition suggests that the larger the deviation of X from XB, the larger the error in estimates of Y. A formula for computing SD is given by:

$$SD = SE\left[1 + \frac{1}{n} + \frac{(X - XB)^2}{n(SX)^2}\right] \tag{29}$$

where n is the sample size to be drawn from Y. Since these simulations usually have large values of n, then for our purposes SD is approximately equal to SE. However, the correct formula is still used in the program for calculating SD (the standard deviation of Y for a given X).

The method described below pertains to sampling from the pair of variables X and Y where X and Y are correlated as above with X independent and Y dependent. First, a value of X is selected at random from the population describing X. This value is then substituted into the regression equation to obtain an estimate of Y which is taken as the expected value or mean of Y for this given value of X. Finally, the standard deviation of Y, SD, is calculated for this value of X. Then a sample is taken from the distribution of Y with the mean given by the estimate of Y from the regression equation and the standard deviation given by SD. This provides a "correlated" random value of Y associated with the given value of X. This procedure is continued until the desired sample size for X and Y is attained.

The following discussion pertains to the type of distribution from which sampling is allowed for Y, the dependent variable in the regression equation. Upon changing from one value of X to the next, the above procedure will simply provide a new mean and standard deviation for Y to be used for sampling purposes. Since the normal and log-normal distributions are the easiest distributions from which to sample when the mean and standard deviation are known, the program allows the user to select only one of these two distributions for sampling from Y. However, the program is designed for expanding this choice to the Weibull and gamma distributions if more subroutines are written and added. Whenever the log-normal distribution is used for Y, the program performs the regression analysis for log Y as a function of log X, so that the sample value for log Y is first calculated and then converted to Y internally by taking antilogarithms. It must be remembered that if one chooses to use the log-normal distribution for Y and to supply R, XB, YB, SX, and SY instead of supplying the raw sample data, then the values of XB, SX, YB, and SY must be given in terms of log X and log Y. That is, SB must be the mean of log X, etc. If the raw data values are supplied for X and Y respectively, the program automatically converts to logs if Y is assigned a log-normal distribution.

SECTION IV

EXAMPLE OF USE OF SIMULATION APPROACH WITH COAL-FIRED ELECTRIC UTILITIES

In order to obtain an indication of how well the simulation procedure approximates the "true" population, SO_2 emissions from the stacks of coal-fired electric utilities were examined. Data were available on % sulfur, coal consumed (CC), and stack height (h) for 224 power plants in the U.S. This was considered to be the total population which was to be simulated by using only a small number (24) of plants in order to obtain information about the distributions of % sulfur, CC, and h.

To obtain a "random" sample, the first 24 plants on the list were selected. Percent sulfur, CC, and h for these 24 plants were then fitted to the four distributions considered in the simulation program. The distributions were then selected which appeared to fit the data better on an overall basis considering the SE, χ^2-value, actual class interval comparisons, and coefficient of skewness and measure of kurtosis calculations. (These techniques of choosing the best fit are discussed in Sections V and VI of this report.) For % sulfur, the Weibull Maximum Likelihood Fit was selected and clipped at the 5% and 99% points. For CC, the Weibull Least Squares Fit was selected and clipped at the 5% and 99% points. Also, h was found not to be independent of CC. Hence, it was decided to treat h as a dependent variable correlated with the independent variable CC by using the raw data on the 24 plants to obtain R, the correlation coefficient. The coefficient of skewness indicated that h was not normal but skewed to the right. Furthermore, the coefficient of skewness and measure of kurtosis for log h indicated "near-normality." Hence, it was decided to use the log-normal distribution for h.

The severity equation for SO_2 emissions from the stacks of coal-fired electric utilities was discussed earlier and is repeated below:

$$S = \frac{(K_2)(\%\ \text{sulfur})(CC)}{h^2} \tag{24}$$

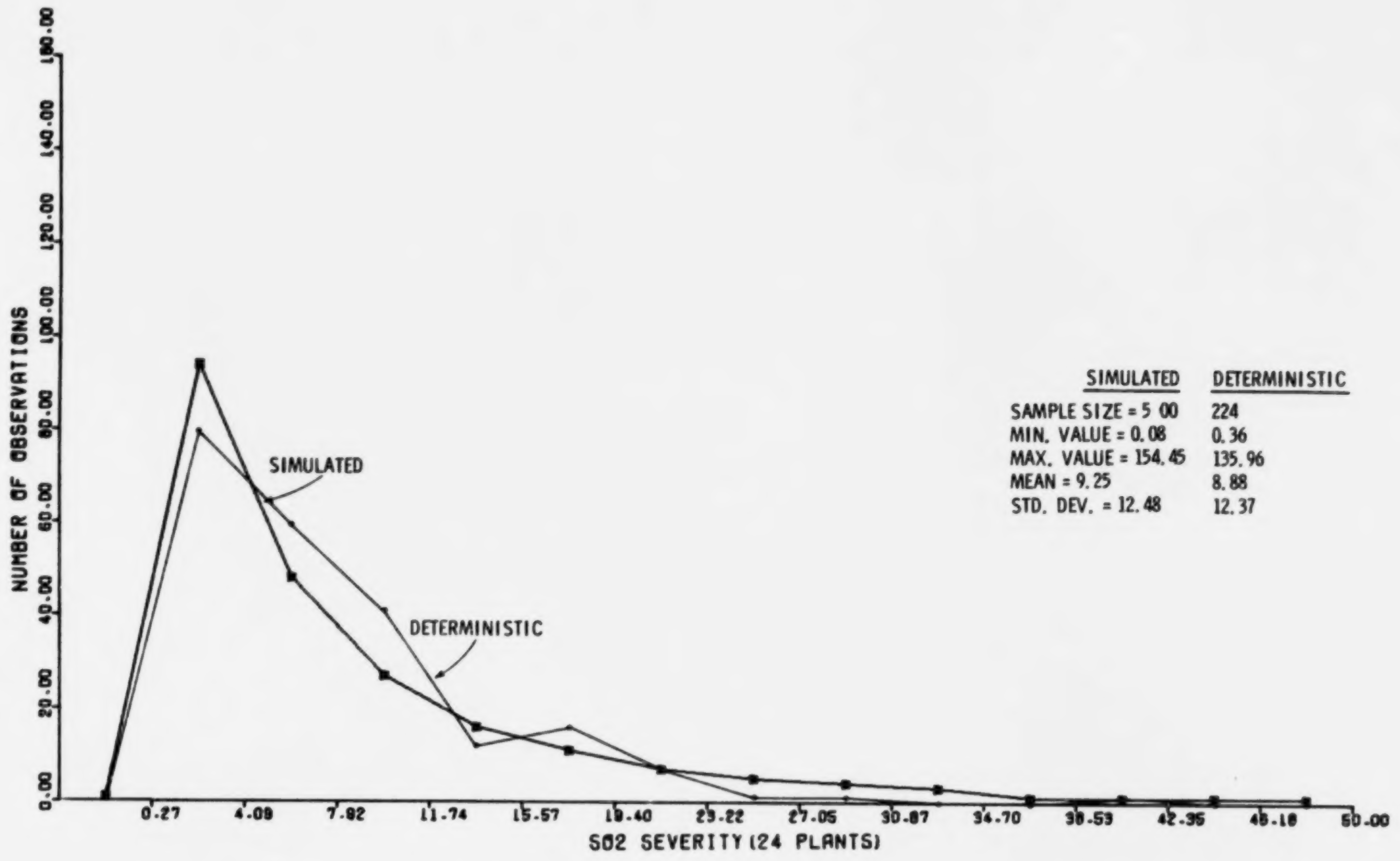
Using this function for S, the data, as indicated above, were entered into the simulation program and 5,000 values were calculated for S. Subsequently, the mean, standard deviation, percentage with S > 1, maximum value, and minimum value were calculated. A deterministic calculation of these values was performed for all 224 plants in the population and the results are compiled in the table below:

Table 2. RESULTS OF DETERMINISTIC CALCULATIONS

Parameter	Simulated value	Deterministic value
Mean	9.25	8.9
Standard deviation	12.5	12.4
Maximum value	154.5	136.0
Minimum value	0.08	0.36
Percent having S > 1	91	95

Frequency histograms and cumulative frequency plots were also drawn for both the simulated values and the deterministic values of S and these are shown in Figures 1 and 2.

The large-sample Z test was performed to determine whether there was a significant difference in the simulated and deterministic mean values obtained above. The test, as



	SIMULATED	DETERMINISTIC
SAMPLE SIZE =	5 00	224
MIN. VALUE =	0.08	0.36
MAX. VALUE =	154.45	135.96
MEAN =	9.25	8.88
STD. DEV. =	12.48	12.37

Figure 1. Frequency histograms for the severity of SO_2 emissions from coal-fired electric utilities comparing the simulation and deterministic methods

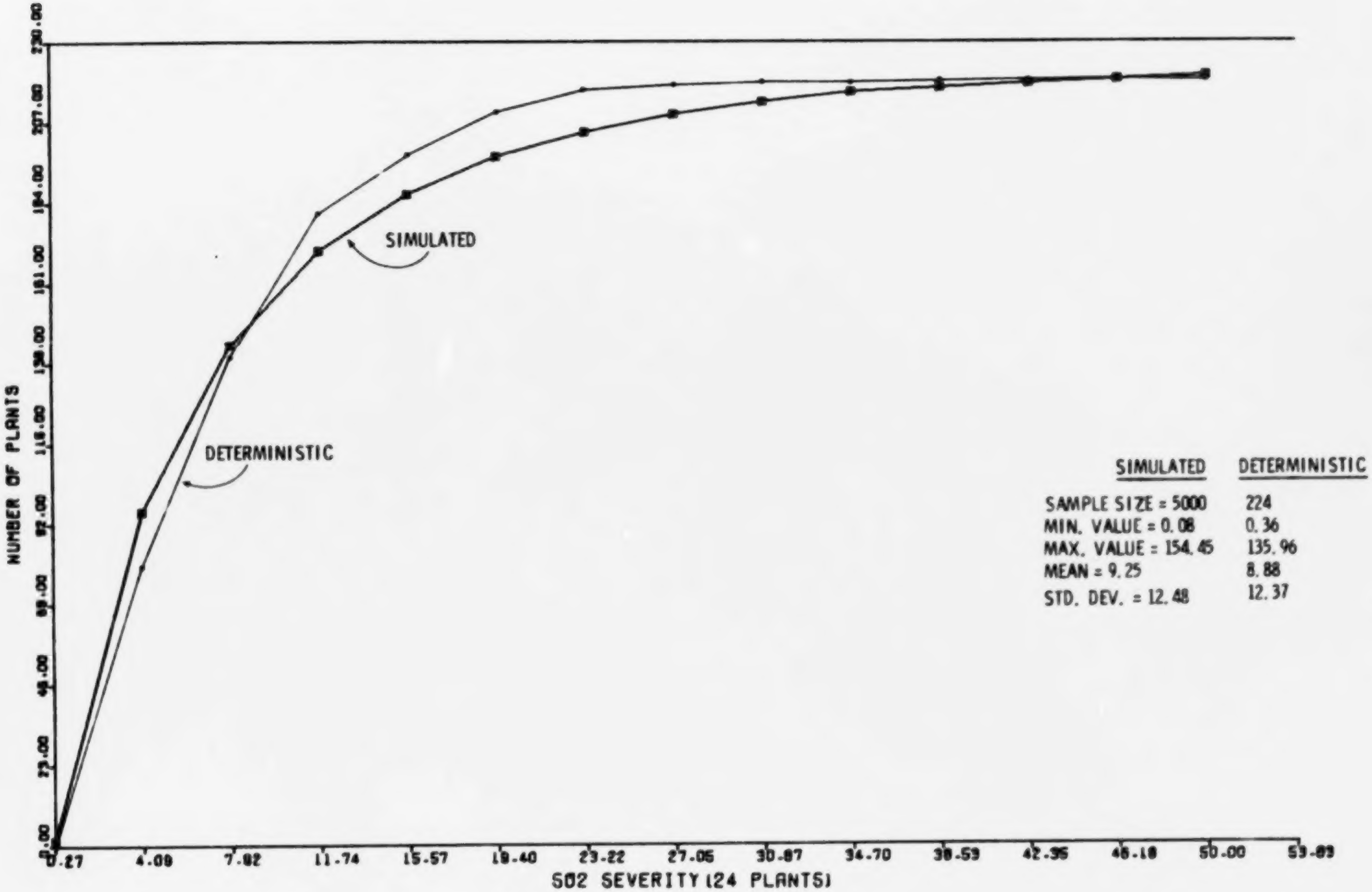


Figure 2. Comparison of cumulative frequency for the severity of SO_2 emissions from coal-fired electric utilities using the simulation and deterministic methods

might be expected, showed no significance in the difference at the 0.01 or 0.05 levels. Furthermore, the F test for significant difference in the variances was also negative, indicating no significant difference.

As can be seen from the frequency histogram of both the simulated and deterministic plots, the severity appears to be log-normally distributed. Furthermore, the cumulative plots for the simulated and deterministic values of S indicate an extreme-value distribution, such as the log-normal distribution. As can be seen by comparing the cumulative plots for the simulated and deterministic values of S, parts of the distributions agree very well whereas other parts do not agree as well. This is considered acceptable since the total population in the deterministic calculation has only 224 points and this is very much a discrete population, whereas the simulation plot assumes a continuous distribution for S. Moreover, the sample correlation coefficient, which was calculated from the 24 plants in the sample, probably does not completely represent the true picture of the actual correlation between CC and h. For these reasons, the χ^2 goodness-of-fit indicates an overall poor fit of the simulated cumulative to the deterministic cumulative distribution.

However, keeping in mind the purpose behind a simulation (viz., to obtain information about an output variable when very little is known about the input variables) and the conditions under which it is generally performed (viz., little or no information about the true distributions of the input variables), the simulated values of the mean, etc., and the cumulative plots are actually remarkably close to the "true" values.

SECTION V

DISCUSSION OF NON-NORMALITY AND CHI-SQUARE GOODNESS-OF-FIT TEST

There are many real-world situations in which massive amounts of data are collected and analyzed. In many of these cases the data come from a population which is finite. Thus, when statistical methods are employed to analyze the data, the underlying distributions are of the discrete type. However, it is common practice to approximate these discrete distributions with one or more of a large number of continuous distributions. The reason, of course, is that continuous distributions are much easier to handle.

After deciding to use continuous distributions in the analysis of the data, the experimenter must decide which of the many types of distributions available will most closely approximate his data and yield the best results for his purposes. In some situations, past experience leads the experimenter to choose a particular type of distribution. For example, the exponential distribution and one of its generalizations, the gamma distribution, are often used to describe the distribution of arrival and/or departure times in a queueing theory problem. In reliability theory, the exponential distribution and another of its generalizations, the Weibull distribution, are often used to describe the distribution of time-to-first-failure or mean time between failures. In measuring air pollution, the log-normal distribution is often used to describe the distribution of the concentrations of pollutants in the air. However, in many

Preceding page blank

instances the experimenter may have no previous knowledge of how his data "should" be distributed. In these instances it is sometimes desirable simply to use the well-known normal distribution to describe the data. With the assumption of normality, many statistical tests and procedures become trivial to apply. However, when normality is assumed and the "true" distribution is non-normal, error is introduced into the problem. Thus, when an experimenter assumes his data are normally distributed in order to be able to apply statistical tests that require the normal distribution, he is naturally concerned with the extent to which his assumption distorts the desired results. One of the purposes of this section is to shed some light on questions of the above type.

This discussion is mainly concerned with data collected in connection with air pollution analysis and it is frequently the case that these data exhibit extreme value characteristics. This means that the data are not symmetrically located around the mode but instead have extreme values to one side or the other of the mode which causes the mean to be shifted to the right or left of the mode. Thus, the data cannot generally be described by the normal distribution. Analytical techniques that can be used to detect and measure certain "degrees of non-normality" of data will be discussed. Some of the well-known extreme value distributions, e.g., the Weibull, gamma, and log-normal distributions, will also be reviewed and compared with the normal distribution for different values of their parameters. Finally, the χ^2 goodness-of-fit test will be discussed to show how it can be used to test the fit of the data to a given distribution.

A. CENTRAL LIMIT THEOREM AND T-TEST

Consider a given population which has an unknown distribution with a finite mean and variance. Let $X_1, \ldots, X_n$ denote a random sample from this population so that each X_i is a random variable with the same distribution as the underlying population distribution. Table 3 gives the notation for certain statistical parameters for both the sample and the population.

Table 3. NOTATION USED FOR STATISTICAL PARAMETERS

Parameter	Sample notation	Population notation
Mean	$\overline{X}$	μ
Variance	$(SD)^2$	σ^2
Third central moment	m_3	μ_3
Fourth central moment	m_4	μ_4
.	.	.
.	.	.
.	.	.
k^{th} central moment	m_k	μ_k

In general, English letters are used to denote a sample parameter and Greek letters are used to denote the corresponding population parameter. The first result to be discussed is the so-called "law of large numbers." This law states that each of the sample characteristics in the table above (as well as some others to be discussed later) converges in probability to the corresponding population parameter. It is instructive to see what this means for the case of $\overline{X}$, the sample mean. Let $\varepsilon > 0$ be any positive member. Then:

$$\lim_{n\to\infty} [P(|\overline{X}_n-\mu|>\varepsilon)] = 0 \qquad (30)$$

or
$$\lim_{n\to\infty} [P(|\overline{X}_n-\mu|\leq\varepsilon)] = 1 \tag{31}$$

where $\overline{X}_n$ denotes the mean of a sample of size n. This indicates that the probability that the mean of a sample of size n will differ from the true population mean by as much as $\varepsilon > 0$ (for any $\varepsilon > 0$) approaches 0 as the sample size n approaches $+\infty$. The same is true of the other sample characteristics.

Practically speaking, the law of large numbers is of little benefit unless one can determine how good the approximation is as a function of sample size or unless confidence intervals can be found for certain samples of size $n \geq 30$ where the population parameters are "reasonably approximated" by their sample counterparts. Instead of simply taking the sample parameter [e.g., $\overline{X}$ or $(SD)^2$] as the value of the population parameter (μ or σ^2), it is usually desirable instead to construct confidence intervals around the sample parameter within which the target population parameter is assumed to be with a certain degree of confidence. Since an approximation to the population mean and a confidence interval about it seem to be of central importance in most statistical analyses of data, procedures (and their limitations) for accomplishing this task are described below.

First, it is necessary to consider some preliminary definitions and results. A random variable is said to have a chi-square distribution with n degrees of freedom if its p.d.f. is given by:

$$f(X) = \left\{ \begin{array}{ll} \dfrac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)} \; X^{\left(\frac{n}{2}-1\right)} \; e^{-\frac{X}{2}} & \text{for } X > 0 \\ & \\ 0 & \text{elsewhere} \end{array} \right\} \tag{32}$$

Suppose a random sample $X_1, \ldots, X_n$ is available from a normal population with mean, μ, and variance, σ^2. Then the random variable defined by:

$$\chi^2 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})^2}{\sigma^2} = \frac{(n-1)(SD)^2}{\sigma^2} \qquad (33)$$

has a chi-square distribution with n-1 degrees of freedom.

Next, suppose that Z is a standard normal random variable and χ^2 is any chi-square random variable with n degrees of freedom. Then the random variable:

$$T = \frac{Z}{\sqrt{\frac{\chi^2}{n}}} \qquad (34)$$

is defined to be the "Student t" random variable with n degrees of freedom. The T variable is a symmetric distribution that looks very much like the normal except that its tails are somewhat wider.

The T-test is used to find confidence intervals about the mean of a sample. Let $X_1, \ldots, X_n$ denote a sample of any size from a normal distribution with unknown mean, μ, and variance, σ^2. Let $\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$ be the sample mean and let $(SD)^2$ be the (unbiased) sample variance. By standard statistical theorems, $\overline{X}$ is designated as a normal random variable with mean, μ, and variance, $\frac{\sigma^2}{n}$. Also, by one of the results stated above, $\frac{(n-1)(SD)^2}{\sigma^2}$ is found to be a chi-square random variable with n-1 degrees of freedom. Thus:

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\frac{(SD)}{\sigma}} = \frac{\overline{X}-\mu}{SD/\sqrt{n}} \qquad (35)$$

is a "Student t" random variable with n-1 degrees of freedom. Thus, by using a table for the T variable, one can find $t_{\frac{\alpha}{2}}$ such that:

$$P(|T| \leq t_{\frac{\alpha}{2}}) = 1 - \alpha \tag{36}$$

for any given α. For example, if $\alpha = 0.05$, then $t_{0.025} = 2.2$ for 11 degrees or $t_{0.025} = 2.18$ for 12 degrees of freedom, etc. Hence given α and the number of degrees of freedom, a value $t_{\frac{\alpha}{2}}$ can be found such that

$$-t_{\frac{\alpha}{2}} \leq T \leq t_{\frac{\alpha}{2}}$$

with probability $1-\alpha$. Thus, with $\alpha = 0.05$, one finds a 95% confidence interval for T. For the T defined above, it is found that $\left|\frac{X-\mu}{SD/\sqrt{n}}\right| \leq t_{0.025}$ with 95% confidence

$$\Longrightarrow \quad -t_{0.025} \leq \frac{\mu-\overline{X}}{SD/\sqrt{n}} \leq t_{0.025}$$

$$\Longrightarrow \quad \overline{X} - t_{0.025}\frac{SD}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{0.025}\frac{SD}{\sqrt{n}}$$

Thus, an interval is obtained in which the population mean lies with 95% certainty.

The above procedure will now be examined with reference to the assumption of normality. It is useful to establish the reason why it was necessary for the sample to be from a normal population. This question can be answered by reversing the steps used above. In defining the T random variable, the numerator had to be a standard normal random variable and the denominator had to be $\sqrt{\frac{\chi^2}{n}}$ where χ^2 had a chi-square distribution with n degrees of freedom. Thus, if the sample were taken from a non-normal population, the sample mean, $\overline{X} = \frac{\Sigma X_i}{n}$, will not (in general) be a normal random variable and the variable $\frac{(n-1)(SD)^2}{\sigma^2}$ will not in general have a chi-square distribution. Thus, for non-normal populations, the ratio used in defining the T variable will not produce a T variable and, hence, the T-test is not strictly valid. In such cases, the central limit theorem is used. This theorem states that if a sample of size n is taken from any population with finite mean, μ, and variance, σ^2, the sample mean, $\overline{X}_n$, approaches a normal random variable in distribution as n increases. In fact, for samples of size $n \geq 30$, the sample mean is so close to a normal distribution that for all practical purposes it can be considered normal. Thus, if a sample of size $n \geq 30$ is taken from any population and its mean is calculated, one will obtain approximately a normal random variable with mean, μ (the unknown population mean), and variance $\frac{\sigma^2}{n}$ (where σ^2 is the unknown population variance). Hence, letting

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \qquad (37)$$

a standard normal random variable is obtained. Since σ is unknown in the above equation and the sample size is relatively large (≥ 30), σ can be approximated by the sample standard deviation, SD, to obtain

$$Z = \frac{\overline{X}_n - \mu}{SD/\sqrt{n}} \qquad (38)$$

for the standard normal random variable. Hence, the normal probability table may be used to find a value $Z_{\frac{\alpha}{2}}$ such that

$$P(|Z| \leq Z_{\frac{\alpha}{2}}) = 1 - \alpha \qquad (39)$$

and, hence, obtain (as in the T-test)

$$\overline{X}_n - Z_{\frac{\alpha}{2}} \frac{SD}{\sqrt{n}} \leq \mu \leq \overline{X}_n + Z_{\frac{\alpha}{2}} \frac{SD}{\sqrt{n}}$$

with 100(1 - α)% confidence.

If the above two procedures are compared, it is found that the ratio under consideration for finding the values $t_{\frac{\alpha}{2}}$ or $Z_{\frac{\alpha}{2}}$ is the same in both cases, namely, $\frac{\bar{X}-\mu}{SD/\sqrt{n}}$, and the only difference is the specific reference table used. If the underlying population is normal and n is arbitrary, the T table is used and if not, but $n \geq 30$, the normal table is used. Actually, the T tables in most statistics books stop at $n = 30$ (29 degrees of freedom) because for $n \geq 30$ the variable defined by the above ratio can be considered to be normal or T regardless of the underlying population distribution. Hence, it would be useless to construct tables beyond approximately $n = 30$ for the T variable since they would simply duplicate the values that could be obtained from a normal table.

It is sometimes desired to find a confidence interval when n is less than 30 and our data are non-normal. Unfortunately, there is no completely statisfactory analytical answer that can be given in this case. However, statisticians have found from experience that the normal approximations guaranteed by the central limit theorem for samples of size $n \geq 30$ are still very close to normal for most mound-shaped probability distributions. In some cases, the approximation is valid for samples as small as $n = 5$. Hence, in the case of $n < 30$, one can simply calculate the confidence interval in the same way as described above, assuming normality, and subsequently point out the possible pitfalls that could occur if the underlying distribution strays too far from normality (viz., the confidence interval would no longer be valid and more sampling would be required). An alternate approach for $n < 30$, when there is reason to believe that the data are log-normally distributed, is to apply the T-test to the logs of the data and obtain a confidence interval.

It is interesting to apply the above procedure to the data for coal-fired electric utilities. Data are available for 224 power plants on coal consumption, stack heights, percent sulfur in the coal used, and percent ash in the coal used. The first 24 plants on the list were then selected and considered to be a random sample from the population of 224 plants. Using the 24 plants in the sample, the sample mean, standard deviation, and 95% confidence limits on the mean were calculated for each of the above types of data and these were compared with the population parameters for all 224 plants. The results are provided in Table 4.

Table 4. COMPARISON OF RANDOM SAMPLE VALUES AND POPULATION MEAN FOR COAL-FIRED ELECTRIC UTILITIES

	Random sample values			
Type of data	Mean	Standard deviation	95% confidence interval	Population mean
Coal consumed, kg/yr	1,326	1,349	756;1,896	1,089
Stack height, m	93.6	36	78.4;108.8	101.3
Percent sulfur	1.82	1.15	1.33;2.31	2.48
Percent ash	12.25	4.1	10.52;14.0	13.03

As noted above, the 95% confidence interval contains the population mean for each of the data types studied except the percent sulfur. Using various statistical goodness-of-fit tests (to be discussed later) on the data, it was found that coal consumption was exponentially distributed almost perfectly, the percent ash was log-normally distributed, and stack height could not be fitted with much success to any of the distributions that were tried. However, even with these wide-ranging distributions (which are very much non-normal) and a sample size of only 24, the T-test still provided satisfactory results in every case except the percent sulfur used. Upon investigating the percent sulfur situation more closely, it was found that the range on the data in the sample of the first 24 plants was from 0.4% to 4% whereas the population data ranged from 0.4% to 6.2%. This indicates that, in the case of percent sulfur, the random sample is probably not very good. However, the 95% confidence interval for percent sulfur was still relatively close to containing the actual population mean.

B. COEFFICIENT OF SKEWNESS AND KURTOSIS AS ANALYTICAL MEASURES OF NON-NORMAL DISTRIBUTIONS

In a symmetric distribution, every moment of odd order about the mean (if existent) must be zero. Any such moment which is not zero may thus be considered as a measure of the asymmetry or skewness of the distribution. The simplest of these measures is μ_3 which is of the third dimension in units of the variable. In order to reduce this to zero dimension and so construct an absolute measure, division by σ^3 is performed and the ratio

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad (40)$$

is regarded as a measure of the skewness. γ_1 is called the coefficient of skewness.

In statistical applications, unimodal continuous distributions of the type shown in Figure 3 are often encountered. The frequency curve forms a long tail to the right of the mode and a short tail to the left. Thus, in calculating μ_3, the cubes of the positive deviations will generally outweigh the negative cubes and, hence, μ_3 will be positive as will $\gamma_1 = \frac{\mu_3}{\sigma^3}$. Thus, the above distribution is said to be skewed right or have positive skewness. Similarly, negative skewness occurs when $\gamma_1 < 0$.

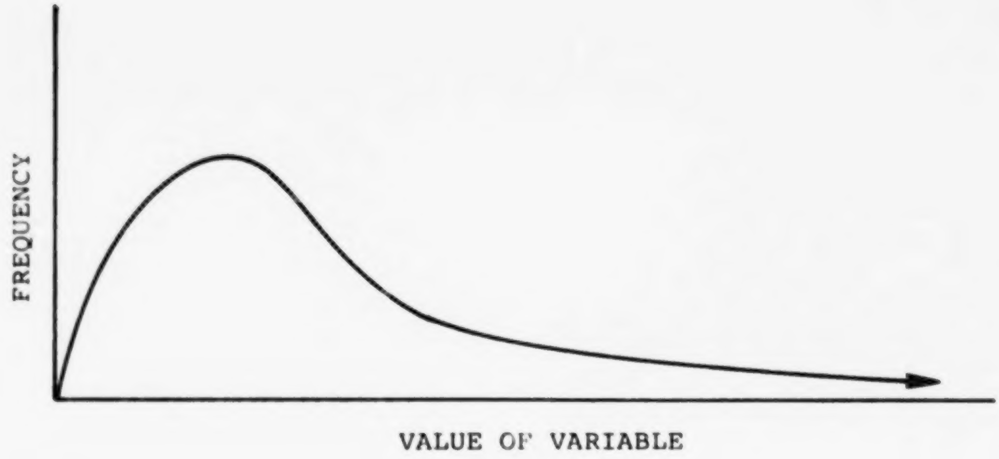


Figure 3. Unimodal continuous distribution

Reducing the fourth moment μ_4 to zero dimension in the same way as above, the measure of kurtosis of a distribution is defined as:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} \tag{41}$$

The measure of kurtosis is used as a measure of the degree of flattening of a frequency curve near its center. The normal distribution has a constant measure of kurtosis, $\gamma_2 = 3$. Thus $\gamma_2 > 3$ means that the distribution has a sharper peak, thinner shoulders, and fatter tails than the normal distribution. Likewise, $\gamma_2 < 3$ means that the distribution has flatter peaks, fatter shoulders, and thinner tails than the normal distribution. Figure 4 exhibits these features.

All of the above distributions are not skewed although curve 1 and curve 2 in Figure 4 would fail to be normal because of deviations in their measures of kurtosis.

It can be shown that $\gamma_1 = 0$ and $\gamma_2 = 3$ for the normal distribution. Hence the values of skewness and kurtosis for a

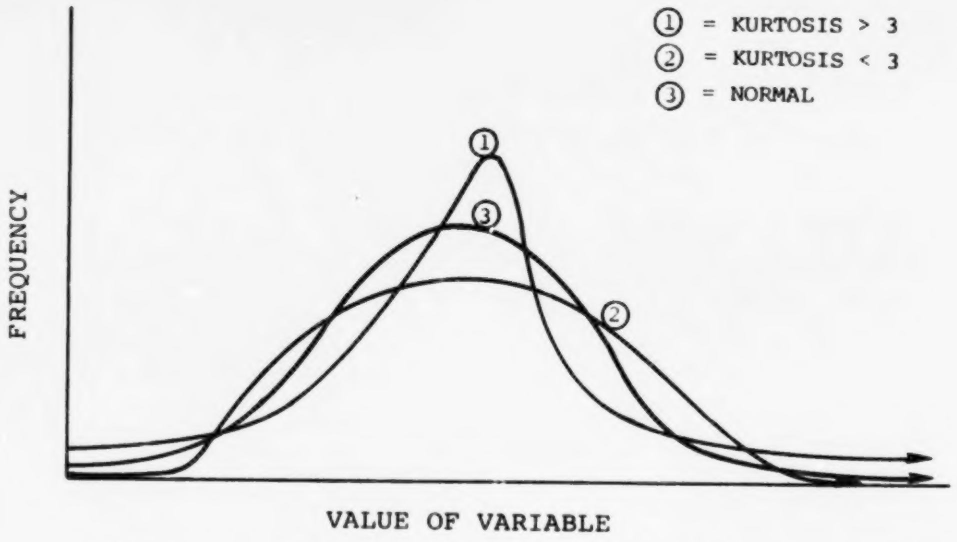


Figure 4. Examples of kurtosis in distributions

given distribution can be used to compare it with the normal. To obtain another reference point for comparisons, consider the exponential distribution function whose probability density is given by:

$$f(X) = \begin{Bmatrix} 1/\beta\ e^{-X/\beta} \text{ for } X > 0 \\ 0 \text{ elsewhere} \end{Bmatrix} \tag{42}$$

where $\beta > 0$ is an arbitrary parameter. By calculating the necessary moments, etc., it is found that the exponential distribution also has constant values for γ_1 and γ_2 independent of the parameter β. These are given by:

$$\begin{aligned} \gamma_1 &= \text{coefficient of skewness} = 2 \\ \gamma_2 &= \text{measure of kurtosis} = 9 \end{aligned} \tag{43}$$

Hence, another reference point is available for comparison purposes. The exponential distribution is a one-tailed distribution which (as indicated above) is heavily skewed right with a sharper peak, thinner shoulders and a fatter tail than that of the corresponding normal distribution. The density function for the exponential distribution is graphed below.

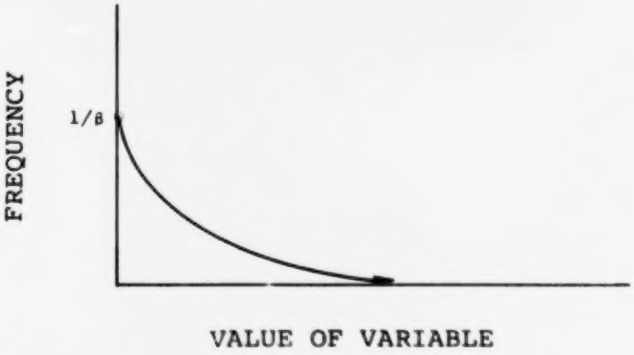


Figure 5. Probability density function for the exponential distribution

In the next sections, the formulae will be investigated for various points of interest on the Weibull and gamma distributions as a function of their parameters. The mean, median, mode, value of the p.d.f. at the mode (i.e., the maximum value of the p.d.f.) and the coefficient of skewness and measure of kurtosis will be considered.

C. WEIBULL DISTRIBUTION

The two-parameter family of Weibull density functions is given by:

$$f(X) = \begin{Bmatrix} abX^{b-1} e^{-aX^b} & \text{for } X > 0 \\ 0 & \text{otherwise} \end{Bmatrix} \tag{44}$$

where a, b > 0 are arbitrary parameters. The cumulative distribution can be found in closed form and is given by:

$$F(X) = \begin{Bmatrix} 1-e^{-aX^b} & \text{for } X > 0 \\ 0 & \text{elsewhere} \end{Bmatrix} \tag{45}$$

Graphs of the p.d.f. of the Weibull distribution for various values of b are given below (for a = 1):

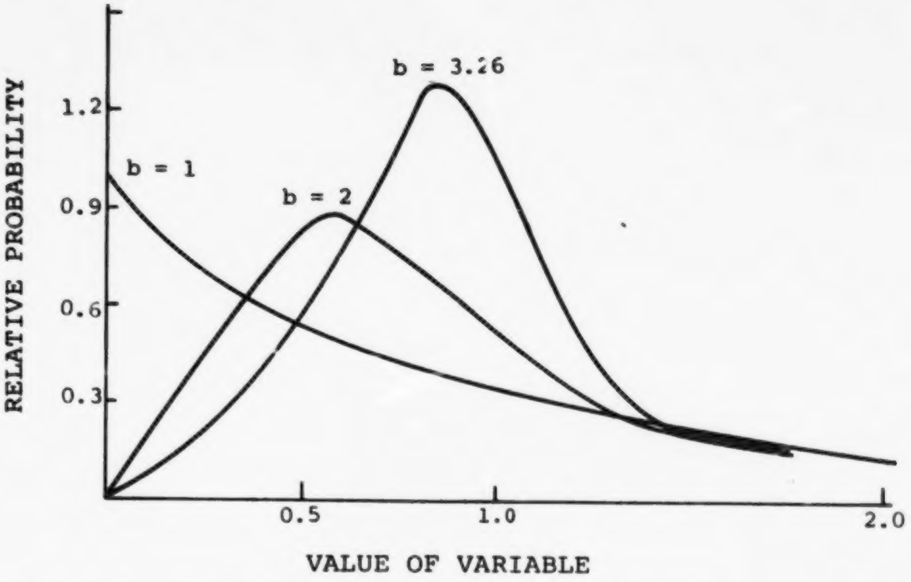


Figure 6. Probability density function of the Weibull distribution for various values of parameter b (for a = 1)

From Figure 6, it can be seen that for b = 3.26 the curve is almost normal; for b = 2, it is skewed right with positive kurtosis; and, for b = 1, it is an exponential curve. For b = 1, it can be seen that the Weibull distribution reduces precisely to the exponential distribution. In order to observe how the curve changes for different values of a and b, some of its points of interest are analytically calculated below as a function of a and b. The following formulae are used:

$$\mu = \text{mean} = a^{-1/b}\Gamma(1+1/b) \tag{46}$$

$$\text{median} = a^{-1/b}(\log 2)^{1/b} \tag{47}$$

$$\text{mode} = a^{-1/b}\left(\frac{b-1}{b}\right)^{1/b} \quad (\text{for } b \geq 1;\ \text{if } b < 1, \text{ the mode} = 0) \tag{48}$$

$$f(\text{mode}) = \text{maximum value of p.d.f.} = \left[ab\left(\frac{b-1}{e}\right)^{b-1}\right]^{1/b} \tag{49}$$

$$(\text{for } b > 1;\ \text{if } b < 1 \text{ the max. value} = +\infty)$$

$$\sigma^2 = \text{variance} = a^{-2/b}[\Gamma(1+2/b) - \Gamma^2(1+1/b)] \tag{50}$$

$$\sigma = \text{standard deviation} = \sqrt{\sigma^2} \tag{51}$$

$$\begin{aligned}\mu_3 &= \text{third central moment} \\ &= a^{-3/b}[\Gamma(1+3/b) - 3\Gamma(1+1/b)\Gamma(1+2/b)+2\Gamma^3(1+1/b)] \\ &= a^{-3/b}\Gamma(1+3/b) - 3\mu\sigma^2-\mu^3\end{aligned} \tag{52}$$

$$\begin{aligned}\mu_4 &= \text{fourth central moment} \\ &= a^{-4/b}[\Gamma(1+4/b) - 4\Gamma(1+1/b)\Gamma(1+3/b) + \\ &\qquad 6\Gamma^2(1+1/b)\Gamma(1+2/b) - 3\Gamma^4(1+1/b)] \\ &= a^{-4/b}\Gamma(1+4/b) - 4\mu\mu_3 - 6\mu^2\sigma^2-\mu^4\end{aligned} \tag{53}$$

$$\gamma_1 = \text{coefficient of skewness} = \frac{\mu_3}{\sigma^3} \tag{54}$$

$$\gamma_2 = \text{measure of kurtosis} = \frac{\mu_4}{\sigma^4} \tag{55}$$

Letting a = 1, the above formulae are used to calculate the points of interest for various values of b as given in Table 5.

As can be seen from Table 5 at b = 3.26, the Weibull distribution function is "almost" normal. For example, its mean, mode, and median all occur at approximately the same point. Furthermore, its coefficient of skewness is near 0 and its measure of kurtosis is near 3 at that value of b. As b decreases from 3.26 to the values of 2.0, 1.5 and lower, the curve becomes more and more non-normal. Its mode begins to shift farther to the left of its mean (and median) indicating a skewed right distribution. The coefficient of skewness (as expected) begins to increase from 0.089 to 0.631, 1.08, etc. Finally, the measure of kurtosis begins to increase also, indicating sharper peaks, thinner shoulders and fatter

Table 5. VALUES FOR VARIOUS POINTS OF INTEREST IN THE WEIBULL DISTRIBUTION FOR VARIOUS VALUES OF PARAMETER B (WHEN A = 1)

Point of interest	Value				
	b=3.26	b=2.0	b=1.5	b=1.4	b=1.1
Mode	0.894	0.707	0.481	0.409	0.113
Value of f(mode)	1.26	0.858	0.745	0.736	0.808
Median	0.894	0.833	0.783	0.770	0.716
Mean (μ)	0.896	0.886	0.903	0.911	0.965
Variance	0.091	0.215	0.375	0.436	0.771
3rd moment about μ	0.002	0.0628	0.248	0.343	1.174
Standard deviation (σ)	0.303	0.463	0.613	0.660	0.878
$F(\mu+\sigma)$	0.836	0.838	0.845	0.848	0.859
$F(\mu-\sigma)$	0.166	0.163	0.145	0.134	0.081
% between $\mu-\sigma$ and $\mu+\sigma$	67%	67.5%	70%	71.4%	77.8%
Coefficient of skewness (= 0 for normal)	0.089	0.631	1.08	1.19	1.73
Measure of kurtosis or excess (= 3 for normal)	2.73	3 244	4.384	4.838	7.296
4th moment about μ	0.023	0.150	0.619	0.918	4.336

tails than the normal. At b = 1, the Weibull distribution reduces to the exponential distribution with $\gamma_1 = 2$ and $\gamma_2 = 9$. It can be seen that γ_1 and γ_2 are converging to these values for when b = 1.1, $\gamma_1 = 1.73$ and $\gamma_2 = 7.3$. If b is allowed to attain values < 1, then as indicated earlier the mode becomes 0 and the maximum value of f becomes +∞. Further, the parameters γ_1 and γ_2 continue to increase giving larger values than those corresponding to the exponential distribution. When b > 3.26, the Weibull distribution becomes skewed left and its kurtosis drops below 3. From the above discussion it can be seen that the parameter b controls the shape of the Weibull distribution, i.e., whether it is "almost" normal or skewed right or left, etc.

Consider what happens whenever the parameter, a, changes for given values. Table 6 is similar to Table 5 except that the value of a has been changed from 1.0 to 1.0 x 10^{-5}. This change in the value of a "stretches" the p.d.f. to cover data that is wide ranging. It also causes the maximum value of the p.d.f. to decrease to the point where the curve is almost flat. The values of γ_1 and γ_2 do not change since they are unitless measures and should not be affected.

D. GAMMA DISTRIBUTION

The gamma distribution is another of the extreme value distributions. The two-parameter family of the gamma p.d.f. is given by:

$$f(X) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} X^{\alpha-1} e^{-X/\beta} & \text{for } X > 0 \\ 0 \text{ otherwise} & \end{cases} \qquad (56)$$

where α, $\beta > 0$ are arbitrary parameters. It can be noted that when $\alpha = 1$, the gamma distribution reduces to the exponential distribution. When $\alpha = n/2$ and $\beta = 2$ for a positive integer n, the gamma distribution reduces to the chi-square distribution with n degrees of freedom.

The closed form for the cumulative distribution function of the gamma distribution does not in general exist and in order to use a table to look up its values one should consult a table of the incomplete gamma function. Of course, if $\alpha = 1$, the closed form exists and, if $\alpha = n/2$ and $\beta = 2$ for a positive integer n, the appropriate chi-square table can be used. The graph of the p.d.f. of the gamma distribution for various values of α with $\beta = 1$ (fixed) is shown in Figure 7.

As can be seen from Figure 7, the gamma distribution is a skewed right distribution with relatively fat tails. The

Table 6. VALUES FOR VARIOUS POINTS OF INTEREST IN THE WEIBULL DISTRIBUTION FOR VARIOUS VALUES OF PARAMETER b (WHEN $a = 1.0 \times 10^{-5}$)

Point of interest	Values				
	b = 3.26	b = 2.0	b = 1.5	b = 1.4	b = 1.1
Mode	30.6	223.6	1,036	1,525	3,968
Value of f(mode)	0.037	0.003	0.0003	0.0002	0.00002
Median	30.6	263.4	1,687	2,870	25,140
Mean (μ)	30.6	280.2	1,946	3,396	33,880
Variance	107.5	2.15×10^4	1.74×10^6	6.06×10^6	9.50×10^8
3rd moment about μ	79.8	1.97×10^6	2.98×10^9	1.78×10^{10}	5.08×10^{13}
Standard deviation (σ)	10.4	146.6	1,319	2,461	30,831
$F(\mu+\sigma)$	0.836	0.838	0.845	0.848	0.859
$F(\mu-\sigma)$	0.165	0.163	0.145	0.134	0.066
% between $\mu-\sigma$ and $\mu+\sigma$	67.1%	67.5%	70%	71.4%	79.3%
Coefficient of skewness[a]	0.089	0.631	1.08	1.19	1.73
Measure of kurtosis[a]	2.73	3.248	4.39	4.82	7.294
4th moment about μ	3.41×10^4	1.5×10^9	1.33×10^{13}	1.77×10^{14}	6.59×10^{18}

[a] The slight variations between the coefficient of skewness and the measure of kurtosis values shown in Tables 5 and 6 are due to calculator round-off, and actually should be identical.

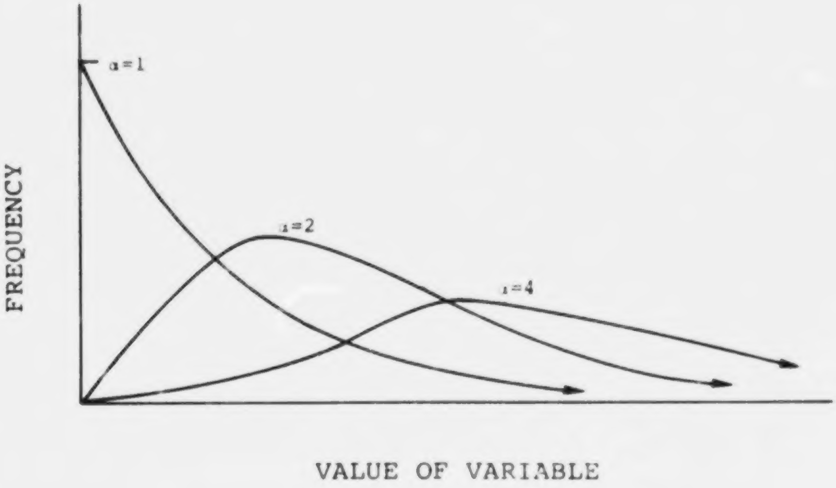


Figure 7. Probability density function for the gamma distribution for various values of parameter α (for $\beta = 1$)

parameter α is the shape parameter and the parameter β is the stretch parameter. To get an analytical picture of skewness, kurtosis, etc., for the gamma distribution, some formulae are provided below for calculating the mean, variance, etc.

$$\mu = \text{mean} = \alpha\beta \tag{57}$$

$$\sigma^2 = \text{variance} = \alpha\beta^2 \tag{58}$$

$$\text{mode} = (\alpha-1)\beta \tag{59}$$

$$f(\text{mode}) = \text{maximum value of p.d.f.} = \frac{(\alpha-1)^{\alpha-1}}{\Gamma(\alpha)\beta}\, e^{-(\alpha-1)} \tag{60}$$

$$\sigma = \text{standard deviation} = \sqrt{\sigma^2} = \beta\sqrt{\alpha} \tag{61}$$

$$\mu_3 = 2\alpha\beta^3 \tag{62}$$

$$\mu_4 = 3\alpha^2\beta^4 + 6\alpha\beta^4 \tag{63}$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{2\alpha\beta^3}{\beta^3\alpha^{3/2}} = \frac{2}{\sqrt{\alpha}} \tag{64}$$

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{3\alpha^2\beta^4 + 6\alpha\beta^4}{\alpha^2\beta^4} = 3 + 6/\alpha \qquad (65)$$

From the above formulae for γ_1 and γ_2, it can be seen that the gamma distribution always has a positive coefficient of skewness and always has a measure of kurtosis > 3. Thus, it will tend to be right skewed and have sharper peaks and fatter tails than the corresponding normal distribution. Only by changing the shape parameter to very large values of α can the gamma distribution be made to approach the normal distribution's values for γ_1 and γ_2. When this is done, the mean and the mode tend to become closer together as they should for the normal distribution. For $\alpha = 1$, the skewness is 2 and kurtosis is 9 as it should be for the exponential distribution of which it is a generalization. Finally, it should be noted that for values of $\alpha < 1$, the skewness and kurtosis increase sharply to very large values.

E. LOG-NORMAL DISTRIBUTION

The two-parameter family of log-normal p.d.f. is given by:

$$f(X) = \begin{Bmatrix} \frac{1}{X\sqrt{2\pi}\beta} \; e^{-1/2\left(\frac{\log X-\alpha}{\beta}\right)^2} & \text{for } X > 0 \\ 0 \text{ elsewhere} & \end{Bmatrix}$$

where α is any real number and $\beta > 0$. A random variable with this p.d.f. has the property that its logarithm to base e is a normal random variable. A graph of the p.d.f. for the log-normal distribution looks similar to the one shown in Figure 8.

As can be seen, the log-normal distribution is another right skewed distribution. Further it tends to have "heavier tails" (i.e., larger kurtosis) than even the exponential

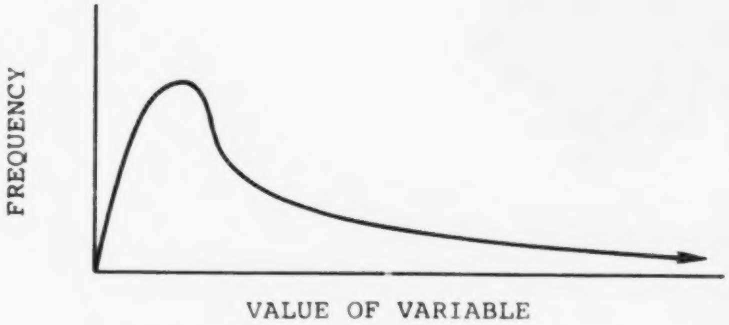


Figure 8. Probability density function for the log-normal distribution

distribution. (This distribution is shown graphically by Curran and Frank.[5])

F. SAMPLE SKEWNESS AND KURTOSIS

Previous discussions were concerned with theoretical distributions and their analytical skewness and kurtosis. It is also instructive to consider samples drawn at random from some population. Consider the sample analogy of skewness and kurtosis and observe how they can be used to predict deviations from normality and toward some other distribution, e.g., the exponential distribution, etc. Formulae are presented below which are used to produce unbiased estimates of the variance and some other higher central moments about the mean:

m_3 = third (unbiased) central moment about $\bar{X}$

$$= \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (X_i - \bar{X})^3 \qquad (66)$$

---

[5] Curran, T. C., and N. H. Frank. Assessing the Validity of the Lognormal Model when Predicting Maximum Air Pollution Concentrations. (Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston. June 15-20, 1975.)

m_4 = fourth (unbiased) central moment about $\overline{X}$

$$= \frac{n^2-2n+3}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} (X_i-\overline{X})^4 - \frac{(2n-3)}{n(n-1)(n-2)(n-3)} \left[\sum_{i=1}^{n} (X_i-\overline{X})^2\right]^2 \qquad (67)$$

$(SD)^2$ = unbiased estimate of variance

$$= \frac{1}{n-1} \sum_{i=1}^{n} (X_i-\overline{X})^2 \qquad (68)$$

From the above, the sample coefficient of skewness, g_1, and measure of kurtosis, g_2, can be constructed as follows:

$$g_1 = \frac{m_3}{(SD)^3} \qquad (69)$$

$$g_2 = \frac{m_4}{(SD)^4} \qquad (70)$$

The law of large numbers discussed earlier can now be enlarged to include the sample parameters g_1 and g_2; i.e., g_1 and g_2 also converge in probability to their population counterparts, γ_1 and γ_2, as the sample size, n, becomes large.

Tables 7 and 8 provide the 0.05 and 0.01 points of the distribution of the coefficient of skewness and the measure of kurtosis, respectively, which can be used to evaluate the sampling distributions for parameters g_1 and (g_2-3). If a given sample has a value of g_1 or (g_2-3) beyond the 0.05 point, it may be deemed to be non-normal. In a stricter test, the 0.01 point would be used. The tables for both g_1 and (g_2-3) are not complete and do not give values for small numbers of samples in the case of (g_2-3). More extensive tables can probably be obtained from the original source shown in the tables. If the calculated values of g_1 and (g_2-3) fall below the 0.05 value, this does not mean that one can categorically accept normality. However, it is a

Table 7. 0.05 AND 0.01 POINTS OF THE DISTRIBUTION OF γ_1, THE COEFFICIENT OF SKEWNESS (NORMAL UNIVERSE)[6,a]

Sample size, n	Probability that γ_1 will exceed listed value in positive direction is	
	0.05 point	0.01 point
25	0.711	1.061
30	0.661	0.982
35	0.621	0.921
40	0.587	0.869
45	0.558	0.825
50	0.533	0.787
60	0.492	0.723
70	0.459	0.673
80	0.432	0.631
90	0.409	0.596
100	0.389	0.567
125	0.350	0.508
150	0.321	0.464
175	0.298	0.430
200	0.280	0.403
250	0.251	0.360
300	0.230	0.329
400	0.200	0.285
500	0.179	0.255
750	0.146	0.208
1,000	0.127	0.180

[a] Points listed are on the positive tail of the distribution; with a minus sign attached, they are equally valid for the negative tail.

---

[6] Geary, R. C., and E. S. Pearson. Tests of Normality. London, University College, 1938.

Table 8. PERCENTAGE POINTS OF THE DISTRIBUTION of γ_2, THE MEASURE OF KURTOSIS (NORMAL UNIVERSE)[6]

Sample size, n	Probability that γ_2 falls below listed value is: 0.01 point	0.05 point	Probability that γ_2 falls above listed value is: 0.05 point	0.01 point
200	-0.63	-0.49	0.57	0.98
250	-0.58	-0.45	0.52	0.87
300	-0.54	-0.41	0.47	0.79
500	-0.43	-0.33	0.37	0.60
1,000	-0.32	-0.24	0.26	0.41
2,000	-0.23	-0.17	0.18	0.28
5,000	-0.15	-0.11	0.12	0.17

good indication that the distribution does not stray far from normality (with respect to skewness and kurtosis) and would not produce serious error in most statistical applications, assuming that it is a normal distribution.

As an example of the above procedure for using g_1 and g_2 and the tables, consider the 224 power plants described earlier and their data on coal consumption, plant capacity, percent sulfur, percent ash, and stack height. Table 9 gives the values of g_1, g_2 and (g_2-3) for these data.

From the conclusions shown in Table 9, it can be seen that, with the exception of sulfur, normality of the sample data can be denied on the basis of skewness and kurtosis alone. Percent ash is denied because of skewness alone and the others fail both tests. The natural question to ask at this point is what distribution (if any) do the data fit if it is definitely non-normal? Since $g_1 = 2.05$ and $g_2 = 8.03$ for coal consumed and $\gamma_1 = 2$ and $\gamma_2 = 9$ for the exponential distribution, one might guess that the coal consumption data are

Table 9. VALUES OF g_1, g_2 AND (g_2-3) FOR POWER PLANT EXAMPLE

Type of data	g_1	g_2	(g_2-3)	Conclusions
Coal consumed	2.05	8.03	5.03	Non-normal
Plant capacity	1.5	4.8	1.8	Non-normal
Stack height	1.46	5.14	2.14	Non-normal
Percent sulfur	0.25	2.88	-0.12	Looks normal as far as skewness and kurtosis are concerned
Percent ash	0.48	3.04	0.04	Non-normal because of slight right skewness; kurtosis is satisfactory

almost exponentially distributed. In the next section, another goodness-of-fit test called the chi-square goodness-of-fit will be discussed. This technique can be used to test how well sample data will fit any variety of distributions and it will be used to analyze the distribution of coal consumed and percent sulfur in the power plants data.

G. THE CHI-SQUARE GOODNESS-OF-FIT TEST

The chi-square test for goodness-of-fit is a general non-parametric statistical test of hypothesis used to test a hypothesis, H_0, of the form: The sample data are distributed according to some given probability distribution. The chi-square test compares a set of actual sample frequencies with a set of frequencies that would be expected on the basis of H_0. If the two sets compare well, the hypothesis is accepted and the data are assumed to be distributed in the way claimed. If the two sets of frequencies compare poorly, the hypothesis is rejected. Since the sampling distribution that is used tends to form a chi-square distribution under the given hypothesis, the test is called a chi-square test.

In order to formulate the chi-square test, let $F_1, \ldots, F_n$ represent the actual frequencies of the sample data in n class intervals. Under the hypothesis (H_0) that the data are distributed according to some given distribution function, let $f_1, \ldots, f_n$ be the theoretical frequencies that would be expected for a sample of the same size from the given distribution. If H_0 is to be true, sample values of the quantity:

$$\chi^2 = \sum_{i=1}^{n} \frac{(F_i - f_i)^2}{f_i} \qquad (71)$$

will tend to form a chi-square distribution. Hence, given a sample and its actual frequency, the theoretical frequencies that would result from H_0 can be calculated and substituted into the above formula to obtain a value of χ^2. Then, by looking at a chi-square table, one can determine whether the sample value of χ^2 is significant enough (at whatever level desired, usually 0.01 or 0.05) to be rejected as a value from the chi-square distribution. If it is, the hypothesis H_0 is rejected and if it is not, H_0 is accepted. In the case of rejection, the probability of error can be stated and one can be as confident as desired in the rejection. However, if rejection cannot be made at a given level of confidence, then H_0 is accepted but this cannot be done with any statement of error concerned.

Before applying the chi-square test, one must consider the number of degrees of freedom to be used for the sampling distribution χ^2. Since the value for χ^2 is extended over n terms (where n = the number of class intervals), it may appear that the sampling distribution has n degrees of freedom. However, some of these degrees of freedom nave been utilized in the construction of the test. The given sample size was used to determine the theoretical frequency of that sample size for each class interval. This reduces the number of degrees of freedom by one to n-1. The "given" distribution in H_0 (the hypothesis to which the test is being applied) is generally one of the many distributions available in statistics and these are generally parametric families of distributions with one, two, or more parameters. In calculating the theoretical frequencies for each class interval from the given distribution, these parameters must be specified in some way. The usual methods of obtaining these parameters for a given sample involve using the data in the sample in some way (e.g., the method of moments, the method of maximum likelihood, or the method of least squares). If the sample data are used to determine the parameters, the number of degrees of freedom must be further reduced by one for each parameter so estimated. Thus, for a two-parameter distribution, the total number of degrees of freedom finally realized for the χ^2 test is n-3 (where n = the number of class intervals used). If the distribution in the hypothesis H_0 has its parameters specified independently of the sample, or if the theoretical frequencies are given without needing to be calculated from a given distribution, it is not necessary to reduce the number of degrees of freedom by any more than one (i.e., to n-1).

Another consideration that arises with the chi-square test which needs some discussion is the problem concerned with the number of theoretical frequencies for each class interval. It is a conservative rule that the theoretical frequencies in any class interval be five or more. If this is not the case, adjacent intervals should be pooled so as to accomplish this task. When two class intervals are pooled, however, the number (n) of class intervals is reduced by one, and, hence, so is the number of degrees of freedom of χ^2. Thus, it is of no benefit to try to use a large number of class intervals to start the test for in the end these will have to be pooled to obey "the rule of 5." Another point to observe is that whenever the number of class intervals has to be reduced to three or less in order to obey the "rule of 5," the number of degrees of freedom for χ^2 drops to zero and, hence, the χ^2 test is not applicable. Thus, for very small sample sizes (approximately 15 or less), the χ^2 test will not be very workable and other measures must be used to get fits to the data.

Actually, the "rule of 5" is conservative and need not be strictly obeyed. For example, if a theoretical frequency for some class interval is around four and the actual frequency is close to that value, it would not be necessary to pool the class intervals and thus reduce the number of degrees of freedom. Furthermore, it has been observed that if an error of 1% can be tolerated in the probabilities read from the chi-square table at the 0.05 level or a 0.2% error at the 0.01 level, then the smallest theoretical frequency can be as low as two if there are at least six degrees of freedom, as low as one if there are 10 degrees of freedom, and as low as 0.5 if there are 25 degrees of freedom. However, all of these low frequencies should occur only once to be allowed; otherwise, pooling must be used according to the "rule of 5" described above.

The data from the power plants will now be used to demonstrate the χ^2 test. Consider coal consumed, which on the basis of skewness and kurtosis was guessed to be an exponential distribution. The method of least squares is used to arrive at values of the parameters a and b in the Weibull distribution from the given data. These values were given by $a = 9.7 \times 10^{-4}$ and $b = 0.9974$. Since $b \sim 1$, the exponential distribution is represented here. A chi-square test will now be performed to see if the data do indeed fit the exponential distribution. Table 10 gives the actual and theoretical frequencies for nine class intervals (determined by Sturge's rule) beginning with the smallest value and proceeding through the largest value.

Table 10. THEORETICAL AND ACTUAL FREQUENCIES FOR NINE CLASS FREQUENCIES (COAL CONSUMED)

Class interval	Theoretical frequency	Actual frequency
1	115.2	116
2	52.0	55
3	27.1	22
4	14.2	14
5	7.4	8
6	3.9	4
7	2.0	3
8	1.1	0
9	1.2	2

In order to obey "the rule of 5," the last four class intervals are pooled into one with a theoretical frequency of 8.2 and an actual frequency of 9. Thus, six class intervals and three degrees of freedom result. The value of χ^2 for this table is 1.3. Referring to a χ^2 table, this value of χ^2 is not significant at the 0.01 level or at the 0.05 level. In fact, it becomes significant at the 0.8 level. This indicates a very good fit and, hence, one can conclude that the data are indeed exponentially distributed.

Recalling that the gamma distribution should also fit the exponential distribution for $\alpha = 1$, a method of moments fit was completed for the gamma distribution to the sample data to obtain the parameters, $\alpha = 1.06$ and $\beta = 1024.8$. Upon doing a χ^2 test, it was found that the number of degrees of freedom was three and $\chi^2 = 1.9$. This value of χ^2 was not significant until the 0.6 level of significance. Hence, a very good fit to the data was also obtained using the gamma distribution.

In view of the above discussion, one would certainly believe that the χ^2 test should reject normality of these data. Using the maximum likelihood fit to the data (i.e., simply using μ = sample mean and σ = sample standard deviation) for the normal, one obtains three degrees of freedom and a χ^2 value of 31.6 which can be rejected at any level for three degrees of freedom.

Since the log-normal distribution is a right skewed distribution, it is reasonable to believe that it may fit these data. Indeed, the χ^2 tests yields six degrees of freedom with $\chi^2 = 3.4$. This value is not significant with this number of degrees of freedom until the 0.75 or 0.8 level. Hence, the log-normal distribution also yields a very good fit to these data.

Next, consider the % sulfur in the coal consumed. Since the skewness and kurtosis tests earlier indicated a possible normal distribution for these data, the normal distribution was fitted to the data by maximum likelihood and a χ^2 test was performed. The table for nine class intervals is shown below:

Table 11. THEORETICAL AND ACTUAL FREQUENCIES FOR NINE CLASS INTERVALS (% SULFUR IN COAL)

Class interval	Theoretical frequency	Actual frequency
1	22.3	32
2	31.4	28
3	46.5	37
4	50.0	51
5	39.0	42
6	22.0	26
7	9.1	4
8	2.7	1
9	1.0	3

A clarification is in order to define the manner in which the last few class intervals should be pooled. Since pooling the last two intervals yields a theoretical frequency of 3.7 with an actual frequency of 4, this will be used to obtain five degrees of freedom and a χ^2 value of 10.3. This value of χ^2 is not significant at the 0.05 level (it becomes significant at about 0.07 or 0.08). Hence, normality for % sulfur is accepted although with some reservations. If the last three intervals had been pooled, four degrees of freedom would have been obtained and χ^2 would be 9.2 which is not significant at 0.05 (it becomes significant also at about 0.07).

Since the fit to the normal distribution for % sulfur is rather poor by χ^2, various fits to other distributions may be evaluated, viz., the Weibull (by both maximum likelihood and least squares) and the gamma and the log-normal distributions. All of these distributions fail, although the Weibull maximum likelihood comes closest with five degrees of freedom and a χ^2 of 18.3. However, this value is significant at 0.01 and consequently the fit is rejected with "99% confidence."

In order to perform chi-square, skewness, and kurtosis tests on sample data, the amount of computation becomes very large. Hence, a computer program was written to perform the computations involved in obtaining these measures for various types of fits to the four distributions which have been considered in this report: the Weibull, gamma, log-normal, and normal distributions. This program and its method of operations are described in Appendix D.

SECTION VI

APPENDIXES

A. Standard Statistical Formulas for Finite Populations

B. Detailed Derivations of the Criteria Pollutant Severity Equations

C. The Simulation Programs

D. The Goodness-Of-Fit Program

E. Treatment of Correlated Data by Linear Transformation

Preceding page blank

APPENDIX A

STANDARD STATISTICAL FORMULAS FOR FINITE POPULATIONS

In certain cases, when sampling from finite populations (e.g., brick kilns, cattle feedlots, etc.) and the sample size, n, is small compared to the total population size, N, corrections must be made to the standard deviation and confidence limit calculations. It should be recognized that there is a difference between the *sample* standard deviation and the estimated *population* standard deviation. When dealing with a population of 400 plants comprising a given source type, N = 400. If 10 plants are surveyed for a stack height, age, capacity, etc., n = 10. Using the data from those ten plants, one can compute the means and standard deviations of the various parameters. But the computed values are the biased mean and standard deviation of the *sample* of 10 points. What is really desired is an unbiased estimate of the mean and standard deviation of the *population* of 400 plants. The symbol "^" will be used over another symbol to stand for an estimate. Assume that a sample mean, $\overline{X}$, and a sample standard deviation, SD, have been computed. An estimate, $\hat{\mu}$, of the total population mean, μ, is:

$$\hat{\mu} = \overline{X} \qquad \text{(A-1)}$$

$$\text{where } \overline{X} = \frac{\Sigma x}{n} \qquad \text{(A-2)}$$

An estimate of the *population* standard deviation, $\hat{\sigma}$, is simply:

$$\hat{\sigma} = SD \sqrt{\frac{n}{n-1}} \sqrt{\frac{N-1}{N}} \qquad \text{(A-3)}$$

$$\text{where SD = sample standard deviation} = \frac{1}{n} \sqrt{n \Sigma x^2 - (\Sigma x)^2} \qquad \text{(A-4)}$$

Earlier it was indicated that the sample standard deviation gave a biased estimate of the corresponding population parameter. The first square root factor in Equation A-3 corrects for this bias. As the sample size, n, becomes larger, the factor approaches unity; hence, the bias is less for larger sample sizes than for small ones. The following table shows this tendency:

Table A-1. FIRST SQUARE ROOT FACTOR IN EQUATION A-3 AS A FUNCTION OF SAMPLE SIZE

Sample size (n)	$\sqrt{\frac{n}{n-1}}$
2	1.4142
5	1.1180
10	1.0541
50	1.0102
100	1.0050
1,000	1.0001

The second factor, $\sqrt{\frac{N-1}{N}}$, in Equation A-3 adjusts for finite population size, and, if n = N, the correction factor is unity or,

$$\hat{\sigma} = SD \qquad \text{(A-5)}$$

which is logical since the sample size is equal to the population size. There is one more important parameter used in calculating confidence limits and that is the estimated standard error of the mean, $\hat{\sigma}_{\overline{x}}$:

$$\hat{\sigma}_{\overline{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \qquad \text{(A-6)}$$

Confidence limits on $\hat{\mu}$ are thus:

$$\hat{\mu} \pm K\hat{\sigma}_{\overline{x}} \qquad \text{(A-7)}$$

where K is the standard "Student t" variable with (n-1) degrees of freedom, and unless otherwise specified should be for the 95% ($\alpha = 0.05$) confidence level.

The preceding formulae are summarized for reader convenience below:

Quantity	Formula	
Sample mean, $\bar{X}$:	$\frac{\Sigma X}{n}$	(A-2)
Sample standard deviation, SD:	$\frac{\sqrt{n\Sigma x^2-(\Sigma x)^2}}{n}$	(A-4)
Estimated population mean, $\hat{\mu}$:	$\bar{X}$	(A-1)
Estimated population standard deviation, $\hat{\sigma}$:	$SD\sqrt{\frac{n}{n-1}}\sqrt{\frac{N-1}{N}}$	(A-3)
Estimated standard error of mean, $\hat{\sigma}_{\bar{x}}$:	$\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	(A-6)

APPENDIX B

DETAILED DERIVATIONS OF THE CRITERIA POLLUTANT SEVERITY EQUATIONS

1. CO SEVERITY EQUATION

Since the primary standard for carbon monoxide, CO, is for a 1-hour averaging time, t = 60 minutes and t_0 = 3 minutes.

Given $$\chi_{max} = \frac{2\,Q}{\pi e u h^2} \qquad \text{(B-1)}$$

Correcting for averaging time:

$$\bar{\chi}_{max} = \chi_{max}\left(\frac{3}{60}\right)^{0.17} \qquad \text{(B-2)}$$

$$= \frac{2\,Q}{\pi e u h^2}\left(\frac{3}{60}\right)^{0.17}$$

$$= \frac{2\,Q(0.6)}{\pi e u h^2}$$

$$= \frac{1.2\,Q}{(3.14)(2.72)(4.5)h^2}$$

or $$\bar{\chi}_{max} = \frac{(3.12 \times 10^{-2})Q}{h^2} \qquad \text{(B-2)}$$

Given $$S = \frac{\bar{\chi}_{max}}{F} \qquad \text{(B-3)}$$

For the criteria pollutants, F is set equal to the primary standard which is 0.04 g/m^3 for CO.

Then, $$S = \frac{\bar{\chi}_{max}}{F} = \frac{(3.12 \times 10^{-2})Qh^{-2}}{0.04}$$

and $$S_{CO} = \frac{0.78\,Q}{h^2} \qquad \text{(B-4)}$$

2. HYDROCARBON SEVERITY EQUATION

The primary standard for hydrocarbons is for a 3-hour averaging time. Thus, $t = 180$ minutes and $t_0 = 3$ minutes.

Given

$$\chi_{max} = \frac{2\ Q}{\pi e u h^2} = \frac{0.052\ Q}{h^2} \quad (B-1)$$

$$\bar{\chi}_{max} = \chi_{max}\left(\frac{3}{180}\right)^{0.17}$$

$$= 0.5\ \chi_{max}$$

$$= \frac{(0.5)(0.052)Q}{h^2} = \frac{0.0260\ Q}{h^2} \quad (B-5)$$

For hydrocarbons, $F_{HC} = 1.6 \times 10^{-4}\ g/m^3$

Then

$$S = \frac{\bar{\chi}_{max}}{F} = \frac{0.026\ Qh^{-2}}{1.6 \times 10^{-4}} \quad (B-6)$$

and

$$S_{HC} = \frac{162.5\ Q}{h^2} \quad (B-7)$$

3. PARTICULATE SEVERITY EQUATION

The primary standard for particulate is for a 24-hour averaging time.

$$\bar{\chi}_{max} = \chi_{max}\left(\frac{3}{1440}\right)^{0.17}$$

$$= \frac{(0.052)Q(0.35)}{h^2} = \frac{(0.0182)Q}{h^2} \quad (B-8)$$

For particulates, $F_p = 2.6 \times 10^{-4}\ g/m^3$

$$S = \frac{\bar{\chi}_{max}}{F} = \frac{0.0182\ Qh^{-2}}{2.6 \times 10^{-4}} \quad (B-9)$$

and

$$S_p = \frac{70\ Q}{h^2} \quad (B-10)$$

4. SO_x SEVERITY EQUATION

The primary standard for SO_x is for a 24-hour averaging time.

Thus,
$$\bar{\chi}_{max} = \frac{(0.0182)Q}{h^2} \qquad (B-11)$$

The primary standard is 3.65×10^{-4} g/m^3.

$$F_{SO_x} = 3.65 \times 10^{-4}$$

and
$$S = \frac{\bar{\chi}_{max}}{F} = \frac{(0.0182)Qh^{-2}}{3.65 \times 10^{-4}} \qquad (B-12)$$

and
$$S_{SO_x} = \frac{50\ Q}{h^2} \qquad (B-13)$$

5. SEVERITY EQUATION FOR NO_x

Since NO_x has a primary standard with a 1-year averaging time, the χ_{max} correction equation cannot be used. Instead the following equation is used:[a]

$$\bar{\chi} = \frac{2.03\ Q}{\sigma_z uX} \exp\left[-\frac{1}{2}\left(\frac{h}{\sigma_z}\right)^2\right] \qquad (B-14)$$

A difficulty arises, however, because a distance, X, from emission point to receptor, is included. To overcome this, the following rationale is proposed:

The equation
$$\chi_{max} = \frac{2\ Q}{\pi e u h^2} \qquad (B-1)$$

is valid for neutral conditions or when $\sigma_z = \sigma_y$. This maximum occurs when

$$h = \sqrt{2}\sigma_z \qquad (B-15)$$

and, since
$$\sigma_z = aX^b \qquad (B-16)$$

---

[a] Personal communication. Bruce Turner, Environmental Protection Agency.

then the distance X_{max} where the maximum concentration occurs is,

$$X_{max} = \left(\frac{h}{\sqrt{2a}}\right)^{\frac{1}{b}} \qquad \text{(B-17)}$$

For neutral conditions, a = 0.113 and b = 0.911.

The following sample calculations illustrate the concentration estimates.

Assume Q = 10 g/s and h = 50 m, then

$$X_{max} = \left(\frac{50}{0.16}\right)^{1.098} = 548.7 \text{ m} \qquad \text{(B-18)}$$

and $\sigma_z = (0.113)(548.7)^{0.911} = 35.4$ m

Assume u = 4.5 m/s, then

$$\begin{aligned}\bar{\chi} &= \frac{2.03\ Q}{\sigma_z u X_{max}} \exp\left[-\frac{1}{2}\left(\frac{h}{\sigma_z}\right)^2\right] \\ &= \frac{(2.03)(10)}{(35.4)(4.5)(548.7)} \exp\left[-0.5\left(\frac{50}{35.4}\right)^2\right] \\ &= (2.32 \times 10^{-4})(0.369)\end{aligned}$$

or $\bar{\chi} = 8.57 \times 10^{-5}$ g/m^3 (B-19)

Simplifying Equation B-14, since $\sigma_z = 0.113\ X^{0.911}$ m and u = 4.5 m,

$$\bar{\chi} = \frac{4\ Q}{X_{max}^{1.911}} \exp\left[-\frac{1}{2}\left(\frac{h}{\sigma_z}\right)^2\right] \qquad \text{(B-20)}$$

$$\begin{aligned}X_{max} &= \left(\frac{h}{0.16}\right)^{1.098} \\ &= 7.5\ h^{1.098} \qquad \text{(B-21)}\end{aligned}$$

and $$\frac{4\ Q}{X^{1.911}} = \frac{4\ Q}{(7.5\ h^{1.098})^{1.911}}$$

$$\bar{\chi} = \frac{0.085\ Q}{h^{2.1}} \ \exp\left[-\frac{1}{2}\left(\frac{h}{\sigma_z}\right)^2\right]$$

$$\sigma_z = 0.113\ X^{0.911}$$

$$= 0.113[(7.5)h^{1.1}]^{0.911}$$

$$= 0.71\ h$$

Therefore,

$$\bar{\chi} = \frac{0.085\ Q}{h^{2.1}} \ \exp\left[-\frac{1}{2}\left(\frac{h}{0.71\ h}\right)^2\right] = \frac{0.085\ Q}{h^{2.1}}\ (0.371)$$

$$\bar{\chi}_{NO_x} = \frac{3.15 \times 10^{-2}\ Q}{h^{2.1}} \qquad \text{(B-22)}$$

Substituting Q and h:

$$\bar{\chi} = 8.5 \times 10^{-5}\ g/m^3$$

Therefore:

$$\bar{\chi} = \frac{3.15 \times 10^{-2}\ Q}{h^{2.1}} = \frac{2.03\ Q}{\sigma_z u X_{max}} \ \exp\left[-\frac{1}{2}\left(\frac{h}{\sigma_z}\right)^2\right] \qquad \text{(B-23)}$$

The NO_x standard is $1.0 \times 10^{-4}\ g/m^3$. Therefore,

$$F = 1 \times 10^{-4}\ g/m^3$$

and the NO_x severity equation is:

$$S_{NO_x} = \frac{(3.15 \times 10^{-2})Qh^{-2.1}}{1 \times 10^{-4}}$$

$$= \frac{315\ Q}{h^{2.1}} \qquad \text{(B-24)}$$

APPENDIX C

THE SIMULATION PROGRAMS

1. THE INPUT TO THE PROGRAM

In this section, details regarding the input to the simulation program are discussed. The input is divided into nine groups, some of which are always required and some of which are optional. Each of these groups is discussed in the order in which it should appear in practice.

a. Input Groups

Group 1: This is a required group and should appear first. It is a single card containing three pieces of data: a title and two flags to indicate later options. The format is 20A2, 2I5.

(ITIL(I),I=1,20) - a title which will be printed as such on output and which will appear below the x-axis on plots. This is read in columns[a] 1-40.

LT - an integer in cols 41-45. Suppose it has been decided that there is a correlation between a pair of input variables such that a simple regression must be performed to obtain the regression line and SE for sampling purposes as described in a previous section. Then as mentioned earlier, the user has the option of either supplying the program with the raw data and having the regression analysis performed on these data or supplying R, XB, SX, YB, and SY directly and using these directly to obtain the regression line information. By the use of the flag LT, the user signals the program which option should be exercised. If LT = 0 (i.e., cols 41-45 are left blank), the program will expect raw data upon which to perform the regression analysis. If LT = 1 (actually any integer other than 0), the program will expect to read in R etc. to perform the regression analysis. In either case, data must be read into the program. However, these data will not be placed as the second group but rather will appear in Group 8 to be discussed below.

---

[a] Columns are subsequently abbreviated as: cols.

The option is available to run the program with no dependent variables. In this case LT = 0 and the number of independent variables NDVAR = 0. NDVAR is in Group 3.

NCFLAG - an integer in cols 46-50. This flag provides the option of using Sturge's rule to set up the number of class intervals and using a value of XMIN and XMAX obtained from the first 50 or so values of the output variables to establish W and the resulting class intervals. If NCFLAG = 0 (cols 46-50 are blank), Sturge's rule and XMIN and XMAX as described earlier will be used as indicated above. If NCFLAG = 1 (anything ≠ 0) the user can read into the program the number, NINT, of class intervals he wishes to use and the value XMIN of the left endpoint of the first class interval and the width, W, of each succeeding class interval. If the user decides to read in NINT etc., this information will be on a card in Group 9 of the input.

Group 2: This is a required group and consists of a single card with three pieces of data: XSAMP, XPØP, and NSAMP. The format is 2F10.3, I5. This program was originally designed to simulate the severity, S, of air pollution concentrations. In doing so, it is sometimes desired to know how many plants from the sample and from the population have predicted severity below 0.1, between 0.1 and 1, and above 1. Thus, the program automatically monitors the number of simulated values of S that fall in these ranges. It then divides the number in each range by the simulated sample size (usually 5,000 or more) to obtain the relative frequency of each range. It then multiplies these relative frequencies by the actual sample size, XSAMP, and population size, XP0P, to give the required predictions for each range. Thus, the program requires XSAMP and XP0P in real format and this is the place that it is provided. Also, the sample size, NSAMP, is entered in integer form to be used (if required) for reading in the raw data for the regression analysis later.

Of course, if the distributions from samples have not been estimated, or if a person is simply not interested in the values between 0.1 and 1, etc., then XSAMP and XP0P can be left blank (and also NSAMP if a regression analysis on raw data is not being performed). However, if these values are left blank, the blank card must still be supplied as Group 2.

Group 3: This is a required group and consists of a single card containing four pieces of information, all integers. The format is 4I5. Due to restricted core size, samples cannot be drawn as large as required (usually 5,000 or more) for each input variable in order that all of these can be used at one time to calculate values of the output variable. Instead, small samples must be drawn out for each input variable and that many values of the output variable must be calculated. This is followed by the statistical manipulations desired [for example keeping a running ΣS and $\Sigma(S)^2$ for later use in finding the mean and standard deviation of output variables] and a repetition of the process as many times as needed to generate the desired sample size. The parameters on this card dictate the manner in which the program does this.

NGRØUP - an integer in cols 1-5, which tells the program the sample size for each pass. This number must be $\leq$ 50.

NPASS - an integer in cols 6-10 which tells the program how many passes to make, drawing out samples of size NGRØUP at each pass. Clearly, the final sample size will be NGRØUP*NPASS.

NIVAR - an integer in cols 11-15. This parameter tells the program how many of the input variables are independent. For example, in a given equation there may be five input variables. The user may decide that one of the variables (e.g., Y) is dependent on (or correlated with) another variable (e.g., X) and that the other three variables are independent. Then NIVAR in this case would be taken to be four: the three that are given independent and the independent variable X in the correlation or regression equation. Thus, there would be one dependent variable, Y, the variable that will be dependent in the regression equation.

NDVAR - an integer in cols 16-20 which gives the number of dependent input variables. In the example above, NDVAR would be one. However, in general, two or more pairs of correlated variables may be present in the equation so that NDVAR could be two or more. The program does not allow a variable to be independent on one pair for correlation purposes and dependent in another. However, the same independent variable can be used for two or more dependent variables and by proper choice this is all that one should ever need.

A word of caution is in order at this point. If NDVAR is $\geq$ 2, a regression analysis must be performed for each correlated pair in order to obtain the regression equation to be used for sampling purposes. It was pointed out earlier that the flag LT in the first data group gave the user the option of either reading in raw data or supplying R etc., directly. The user is cautioned here that if two or more regression fits are necessary, they must both be performed in the same manner, either both with raw data or both with user supplied values of R etc. The program uses LT to get into one of two loops: either read raw data for X_1 and Y_1, perform a regression analysis and then repeat for X_2 and Y_2 etc. until the number of dependent variables is exhausted; or, read R etc. for X_1 and Y_1, perform a regression analysis and then read R etc. for X_2 and Y_2 etc. until the number of dependent variables is exhausted. Finally, it should be noted that NIVAR + NDVAR (= the total number of variables) must be $\leq$ 10.

Group 4: This is a required group and consists of a single card containing codes which tell what type of distribution is to be used for sampling from the independent variables. The values on the card are integers punched in 16I5 format (actually all 16 will not be used). The values on the card are read into an integer array ICØDE(I) from I = 1 to NIVAR. Thus, ICØDE(1) is the code which tells what type of distribution independent variable 1 has, etc. The distributions and their corresponding codes are listed below:

code 1 = Weibull distribution
code 2 = normal distribution
code 3 = gamma distribution
code 4 = log-normal distribution

For example, suppose three independent variables exist and they are ordered as VAR_1, VAR_2, VAR_3. Suppose VAR_1 has a Weibull distribution while VAR_2 has a normal and VAR_3 a log-normal distribution. Then the data card would look like the line below:

cols	1-5	6-10	11-15
	1	2	4

In setting up the function card (to be discussed later), one must exercise caution to be certain that the variables so ordered above will appear in their proper places in the function.

Group 5: This group of data may consist of more than one card depending on the number of independent variables contained in the function. This is the place where the program is given the parameters that go along with the distribution selected for each input (independent) variable. For each distribution, the parameters are listed which the program needs for it as shown below:

Distribution	Program parameters
Weibull	- A and B
normal	- μ and σ
gamma	- α and β
log-normal	- μ and σ for log X or equivalently the log of geometric μ and geometric σ for X.

These parameters are read into a matrix PAR(I,J) which is a dimensional 10 x 2. Thus, each row of the matrix has two components and the rows correspond to the given variables. The parameters for independent variable 1 should be read into the first row; independent variable 2 into the second row, and so on until the independent variables are exhausted. Thus, the program reads values into PAR(I,J) row-wise, i.e., it reads parameters for independent variable 1 into row 1 as the first two values on the data card, etc. The parameters are expected in the same order as listed in the table above, i.e., for the Weibull distribution, A first then B etc. The format is 8E10.3.

As an example, consider the situation described above where three independent variables (VAR_1, VAR_2, and VAR_3) were present and these were Weibull, normal, and log-normal distributions, respectively. The data card for the parameters should contain A then B for the Weibull as the first two entries; then μ then σ for the normal as the next two entries; and, finally, μ then σ for log VAR_3 (or the log of geometric μ then geometric σ for VAR_3) as the last two entries. Thus, there would be six fields (of the eight total) taken up on the card and it might look as that shown below:

cols	1-10	11-20	21-30	31-40	41-50	51-60
	A	B	μ	σ	μ	σ

where the last μ and σ are for log VAR_3.

Group 6: This group of data may consist of more than one card depending on the number of independent variables. Sometimes in fitting continuous distributions to raw data or real-world situations, the continuous distributions tend to take on extremely high or extremely low values which are unrealistic for the given data. This is a particular problem with extreme-value distributions like the Weibull, gamma, or log-normal distributions. Hence, it is desired to "clip" the continuous distributions at appropriate points to keep these unusually large or small values from occurring. This is the data group in which the program is instructed at which points to clip the (independent) distributions. The lower and upper clip for each variable is read into a two-dimensional array CLIP(I,J) which is dimensioned 10 x 2. The procedure of reading one row containing a low clip and high clip, respectively for each independent variable in the order in which they are coded in ICØDE, is the same as that used for the parameters.

The following discussion pertains to the means by which the distributions are clipped. The direct approach to sampling from the Weibull distribution uses the cumulative distribution function to obtain a sample value for X from a given random number, R. Hence, to clip the distribution so that it does not allow values of X below some given value or above some other given value, all that is needed is to find at which points, C_1 and C_2, between 0 and 1, these low and high values occur on the distribution. The program is then supplied with these points, C_1 and C_2, and it will automatically clip the random number generator so that the random numbers generated will be between C_1 and C_2; hence, the corresponding X-value will be in the correct range also. For example, suppose it is desired to exclude values of X below 10 or above 6,000 and that these points occur at $C_1 = 0.02$ and $C_2 = 0.96$, respectively, on the cumulative distribution for the Weibull. The program is then supplied with these parameters (viz., 0.02 and 0.96) as the values for the array CLIP(I,J) and the random number generator is clipped according to the equation below:

$$R = (0.96 - 0.02)R + 0.02 = 0.94R + 0.02$$

This equation will automatically be set up internally.

Next, the problem of clipping the normal and log-normal distributions is discussed. The direct approach to sampling from these distributions uses the probability density function and not the cumulative distribution function. Hence, the clips in these cases cannot be applied in the same manner. Thus, a very direct approach was used for clipping the values of the variable X in this case. The program was supplied with the actual lowest value that X was allowed to assume and likewise the highest value that X was allowed to assume. The program then samples from normal with no restrictions. If the sample value obtained is between the low and high value supplied, it is retained. Otherwise, the program samples again and continues to do so until it obtains a value in the proper range. Then the program continues the above procedure until the desired sample size is attained. A word of caution is in order. In supplying clips for the normal, low and high values for X are supplied directly. However, in supplying clips for the log-normal, low and high values for log X, not X, should be supplied.

If any or all of the variables are to be unclipped, this card(s) must still be supplied. To leave the Weibull distribution unclipped, values of 0.0 and 1.0, respectively, are supplied for the low and high clip. To leave the normal or log-normal distribution unclipped, extremely low or extremely high values (for example, 1.0×10^{-15} and 1.0×10^{15}) are supplied for the low and high clip.

Group 7: This data group may consist of more than one card depending on the number, NDVAR, of dependent variables. If NDVAR is > 1, the program in this data group is given two pieces of information about each dependent variable: (1) which independent variable it is correlated with (i.e. the number of the independent variable, e.g., 1 for VAR_1 etc.); and (2) the type of distribution to be used in sampling for values of the dependent variable (either normal = 2 or log-normal = 4, at present structure). These codes for the dependent variables are read into a two-dimensional integer array IDCØDE(I,J) which is a 10 x 2 matrix. Hence, each row of IDCØDE corresponds to a dependent variable in the same order as the dependent variables are entered, i.e., row 1 for variable 1, etc.

For example, suppose there are five input variables, four of which are independent and one of which is dependent. Suppose further that for the numbering used on the independent variables, the dependent variable is correlated with the third independent variable and the log-normal distribution is desired for sampling from the dependent variable. The data card for this group would then appear as follows:

$$\begin{array}{ccc} \text{cols} & \text{1-5} & \text{6-10} \\ & \underbrace{\qquad}_{3} & \underbrace{\qquad}_{4} \end{array}$$

Group 8: This data group is optional and must appear only if NDVAR is $\geq$ 1. It may consist of more than one card. This data group will contain the information necessary to perform the regression analysis between the independent variables and dependent variables. Depending on the value of the flag LT in Group 1, this data group will contain either the raw data for the pair(s) of correlated variables (independent variable then dependent variable) or the values of R, XB, SX, YB, SY (in that order) for the independent variable X and the dependent variable Y. The format on all cards is 8E10.3.

Note that if NDVAR is $\geq$ 2, the raw data must appear as follows:

$$\left\{\begin{array}{ll} \text{Independent} & \text{VAR}_1 \\ \text{Dependent} & \text{VAR}_1 \end{array}\right\}$$

$$\left\{\begin{array}{ll} \text{Independent} & \text{VAR}_2 \\ \text{Dependent} & \text{VAR}_2 \end{array}\right\} \text{(Even if Independent VAR}_2 = \text{Independent VAR}_1\text{)}$$

etc.

Further, if NDVAR is $\geq$ 2 and it is desired to enter the values of R, etc. it must appear as below:

$$\left\{R_1,\ XB_1,\ SX_1,\ YB_1,\ SY_1\right\}$$

$$\left\{R_2,\ XB_2,\ SX_2,\ YB_2,\ SY_2\right\} \text{ etc.}$$

Group 9: This data group is optional and will appear only if NCFLAG ≠ 0 on the first data card. In this group, specific values of NINT (number of class intervals), XMIN (beginning left endpoint) and W (width of class intervals) are read into the program. As mentioned previously, if NCFLAG = 0, the program will establish class intervals of its own and this data group is omitted. The format used for this group is I5,2F10.3.

b. The Function Card

Whenever changes are made from one run of this program to another run, they are generally made from one function (describing an output random variable as a function of input random variables) to another function (with different input and output random variables). Hence, it is necessary to change one card in the program itself, which will be referred to as the function card. It appears as a card in the function subprogram SF with calling argument IVAL and common arguments VAR(I,J) (as well as some other dummy arguments).

VAR(I,J) is a 51 x 10 matrix in which the samples drawn from the distributions of the input variables are stored for each pass from 1 to NPASS. These samples are stored columnwise, i.e., column 1 contains the sample from VAR_1, etc., until all of the independent variables are sampled. In the first column after the NIVAR, the dependent variables are stored until they are exhausted. After storing the values in VAR(I,J) in this way, the NGRØUP values of the output variable for this pass are calculated. In this calculation, the function card, defined above, is used.

On a single pass of the program, the values of the output variable from I = 1 to NGRØUP are calculated. For a given value of I, this information is transferred to the subprogram SF as

IVAL. For this value of IVAL, a value SF of the output variable is calculated according to the rule specified on the function card. Thus, caution should be exercised in constructing the function card so that it reflects the true function of the input variables in the proper order in which they appear in VAR(I,J).

As an example, consider the function

$$Z = \frac{3.1*A*B^2}{C*D+E}$$

(where * denotes multiplication). Suppose variables A, C, D, and E are taken to be independent and B is dependent and correlated with variable E. If the independent variables are numbered as below:

A = 1
C = 2
D = 3
E = 4

then automatically, variable B will be numbered 5 in VAR(I,J), i.e., column 5 will contain the values of E. Thus the SF card would be:

```
SF = (3.1*VAR(IVAL,1)*(VAR(IVAL,5)**2)/
     (VAR(IVAL,2)*VAR(IVAL,3)+VAR(IVAL,4))
```

c. Example of Overall Input Data

Suppose it is desired to simulate the severity, S, of SO_2 emissions from coal-fired electric utilities. The severity equation is given by:

$$S = \frac{(30.1)(\%\ \text{sulfur})(\text{CC})}{h^2}$$

where % sulfur = percent of sulfur in coal used

CC = coal consumed in 10^6 kg/yr

h = stack height in meters

Assume that % sulfur and CC are to be independent and h is to be correlated with coal consumed. Suppose the type of distribution for each input variable has been determined along with the corresponding parameters and clips according to the information in the table below.

Table C-1. VARIABLES, DISTRIBUTIONS, PARAMETERS AND CLIPS FOR COAL-FIRED ELECTRIC UTILITIES EXAMPLE

Variable	Type of variable distribution	Parameters	Clips
% sulfur-1	2 (normal)	$\mu=2.5$ $\sigma=1.1$	0.025 and 1.0
CC-2	1 (Weibull)	$A=9.7\times10^{-4}$ $B=1.0$	0.05 and 0.99
h-3	4 (log-normal)	-[a]	-[a]

[a] Not needed since this variable is dependent and, hence, correlated.

Assume that the sample size from which the parameters were obtained is 224 out of a population of 600. (This is not a situation where a simulation is required but it can serve as an example.) Suppose that the program is instructed to set up its own class intervals and to perform a regression analysis on the raw data for CC and h. Finally, assume that a simulated sample size of 5,000 is desired. Then the input data would appear as shown in Table C-2. In Table C-2, the values that should appear on the card(s) are underlined and the columns in which they should appear are shown in parentheses beside them.

Table C-2. SUMMARY OF INPUT DATA BY GROUPS FOR COAL-FIRED ELECTRIC UTILITIES EXAMPLE

Group number	Input data[a]
I	Titles (1-40) (blank) (41-45) (blank) (46-50)
II	224.0 (1-10) 600.0 (11-20) 224 (21-25)
III	50 (1-5) 100 (6-10) 2 (11-15) 1 (16-20)
IV	2 (1-5) 1 (6-10)
V	2.5 (1-10) 1.1 (11-20) 9.7×10^{-4} (21-30) 1.0 (31-40)
VI	0.025 (1-10) 1.0 (11-20) 0.05 (21-30) 0.99 (31-40)
VII	2 (1-5) 4 (6-10)
VIII	Raw data for CC[b]
	Raw data for h[b]
IX	Omitted

---

[a] Underlined information represents input data which is to be placed in columns designated by parentheses.

[b] Eight per card in format 8E10.3.

2. DESCRIPTION OF OUTPUT

There are two forms of output from the simulation program: (1) printed output, which is divided into three parts; and (2) the frequency histogram and cumulative frequency function of the simulated sample of output values.

a. The Printed Output

The first item that is printed out is the title read in on the first card of the input. The next item printed is the mean and standard deviation of the output variable. The probabilities that the output variable lies in the range < 0.1, between 0.1 and 1, and > 1, respectively, are subsequently printed. These are followed by the number of plants (outcomes) from the sample which are predicted to fall in each of these three ranges, respectively, and then by the same numbers for the population. These are printed as SNUM(1), SNUM(2), SNUM(3), PNUM(1), PNUM(2), and PNUM(3) for sample number in range 1, etc. Finally, a table is printed out which gives the class intervals, the actual frequency of the sample in each class interval, and the cumulative frequency function for the right endpoint of the class interval.

A few special comments are provided regarding the first and last class intervals. The first class interval printed has as its left endpoint the minimun value of the output variable in the whole sample of size NGRØUP*NPASS. Its right endpoint is the minimum value found in the first NGRØUP values of the output variables, for it is after the first pass that the program automatically sets up class intervals. The last class interval printed has its left endpoint equal to the maximum value obtained on the first pass from the first NGRØUP values and its right endpoint is the overall maximum value for the whole sample. If the user supplies his own class intervals and the lowest value supplied is lower than the actual minimum value calculated by the program, the first class interval will then look "backwards." For example, suppose the user supplied 0 as the beginning class interval value and the lowest observed value was 0.03 during the simulation. The first two lines for the first two class intervals would then appear something like the following:

Class interval	Actual frequency	Cumulative frequency
From 0.03 to 0.0	0.0	0.0
From 0.0 to 0+W	?	?

Likewise, the last class interval could appear "strange" in a similar situation. If one wishes to know XMIN and XMAX for the whole sample, these can be found as indicated above.

b. The Plots

The output plots from the simulation program are self-explanatory and will not be further discussed.

Table C-3. COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      BIT TRAN(16)
      DIMENSION AF(35).CF(35).ITIL(20).EMP(3).PROB(3).
     1SNUM(3).PNUM(3).XF(35).CLIP(10.2).KAF(35)
      COMMON ICODE(10).VAR(51.10).PAR(10.2).SFV(51).IDCODE(10.2).
     1COR(10.5).XSR(250).YSR(250)
C           THIS PROGRAM IS DESIGNED TO PERFORM A COMPUTER
C           SIMULATION TO OBTAIN VALUES OF ONE RANDOM VARIABLE
C           DEFINED AS A FUNCTION OF OTHER RANDOM VARIABLES.
C           ALL THE RANDOM VARIABLES ARE ASSUMED TO BE
C           CONTINUOUS AND CAN BE ANY ONE OF FOUR DIFFERENT
C           DISTRIBUTIONS: THE WEIBULL. THE NORMAL. THE GAMMA.
C           OR THE LOG-NORMAL.  FURTHERMORE. THE PROGRAM WILL
C           ACCOUNT FOR CORRELATION BETWEEN CERTAIN PAIRS OF
C           THE INPUT VARIABLES.
      READ(1.100)TRAN.ICOL1.CY1.CY2.XI1.XI2.ANGL
      WRITE(5.900) TRAN.ICOL1.CY1.CY2.XI1.XI2.ANGL
      READ(1.12) (ITIL(I).I=1.20).LT.NCFLAG
      READ(1.13) XSAMP.XPOP.NSAMP
      READ(1.10) NGROUP.NPASS.NIVAR.NOVAR
      SMEAN=0.0
      SSQ=0.0
      XMIN=1.0E18
      XMAX=0.0
      DO 25 I=1.35
   25 AF(I)=0.0
      DO 26 I=1.3
   26 EMP(I)=0.0
      CALL RANDU(9.IY.YFL)
      IX=IY
      READ(1.10)(ICODE(I).I=1.NIVAR)
      READ(1.11)((PAR(I.J).J=1.2).I=1.NIVAR)
   11 FORMAT(8E10.3)
      READ(1.11)((CLIP(I.J).J=1.2).I=1.NIVAR)
      IF(NOVAR.EQ.0) GO TO 8
      READ(1.10)((IDCODE(I.J).J=1.2).I=1.NOVAR)
      IF(LT.EQ.0) GO TO 60
      DO 70 I=1.NOVAR
      IROW=I
      CALL SR(NSAMP.IROW.LT)
   70 CONTINUE
      GO TO 8
   60 DO 40 I=1.NOVAR
      IROW=I
      READ(1.11)(XSR(K).K=1.NSAMP)
      READ(1.11)(YSR(K).K=1.NSAMP)
      KO=IDCODE(I.2)
      IF(KO.EQ.4) GO TO 52
      GO TO 50
   52 DO 51 J=1.NSAMP
      XSR(J)=ALOG(XSR(J))
      YSR(J)=ALOG(YSR(J))
   51 CONTINUE
   50 CALL SR(NSAMP.IROW.LT)
   40 CONTINUE
    8 IPASS=1
   22 DO 3 I=1.NIVAR
      ICOL=I
      KK = ICODE(I)
      GO TO (4.5.6.7). KK
    4 CI=CLIP(I.1)
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      C2=CLIP(I,2)
      CALL WSAMP(NGROUP,IX,ICOL,C1,C2)
      GO TO 3
    5 C1=CLIP(I,1)
      C2=CLIP(I,2)
      CALL NORSAM(NGROUP,IX,ICOL,C1,C2)
      GO TO 3
    6 CALL GAMSAM(NGROUP,IX,ICOL)
      GO TO 3
    7 C1=CLIP(I,1)
      C2=CLIP(I,2)
      CALL LOGSAM(NGROUP,IX,ICOL,C1,C2)
    3 CONTINUE
      IF(NDVAR.EQ.0) GO TO 9
      DO 41 I=1,NDVAR
      IVAR=IDCODE(I,1)
      KD=IDCODE(I,2)
      ICOL = I+NIVAR
      IDVAR=I
      GO TO(42,43,44,45),KD
   42 CONTINUE
      GO TO 41
   43 CALL NCSAMP(NGROUP,NPASS,IX,ICOL,IVAR,IDVAR)
      GO TO 41
   44 CONTINUE
      GO TO 41
   45 CALL LCSAMP(NGROUP,NPASS,IX,ICOL,IVAR,IDVAR)
   41 CONTINUE
 9    CONTINUE
      IF(TRAN(1)) CALL BACKT(NGROUP,ICOL1,CY1,CY2,TRAN,XI1,XI2,ANGL)
      CALL SEVAL(NGROUP,XMIN,XMAX,SMEAN,SSQ)
      IF(IPASS.GT.1) GO TO 20
      NCINT=NGROUP*NPASS
      CALL CINT(NCINT,XMIN,XMAX,NINT,XE,W,NCFLAG)
   20 CALL AFREQ(NGROUP,NINT,XE,AF,EMP)
      IF(IPASS.EQ.NPASS) GO TO 21
      IPASS=IPASS+1
      GO TO 22
   21 FN=FLOAT(NCINT)
      DO 23 I=1,3
      PROB(I)=EMP(I)/FN
      SNUM(I)=XSAMP*PROB(I)
      PNUM(I)=XPOP*PROB(I)
   23 CONTINUE
      SDEV=((FN*SSQ)-(SMEAN*SMEAN))/(FN*(FN-1.0))
      SDEV = SQRT(SDEV)
      SMEAN = SMEAN/FN
      S=0.0
      NCL=NINT+2
      DO 24 I=1,NCL
      S=S+AF(I)/FN
      CF(I)=S
   24 CONTINUE
      CALL OUTPUT(SMEAN,SDEV,PROB,SNUM,PNUM,AF,CF,XSAMP,XPOP,
     1NINT,XE,ITIL,XMIN,XMAX)
      KT=40
      NCINT=IFIX(XPOP)
      CALL CIFLOT(NCINT,SMEAN,SDEV,NINT,W,AF,XMIN,XMAX,CF,ITIL,KT)
      DO 333 I=1,45
  333 KAF(I)=((AF(I)/FN)*XPOP)+.5
```

- -

```
      CALL FOPLOT(NCINT,SMEAN,SDEV,NINT,W,XE,XMIN,XMAX,KAF,ITIL,KT)
   10 FORMAT(16I5)
   12 FORMAT(20A2,2I5)
   13 FORMAT(2F10.3,I5)
 100  FORMAT(16L1,I2,1X,5E10.3)
 500  FORMAT(1H0, 16L1,I2,1X,5E15.7)
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE SR(MM,IROW,LT)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51),IOCODE(10,2),
     1COR(10,5),XSR(250),YSR(250)
C          THIS SUBROUTINE DOES EITHER A SIMPLE REGRESSION ON
C          THE INPUT RAW DATA FOR TWO CORRELATED VARIABLES OR
C          READS IN THE VALUES OF R, XB, SX, YB, AND SY AND
C          CALCULATES THE REGRESSION LINE FROM THESE VALUES.
C          IN EITHER EVENT, THE VALUES OF THE INTERCEPT A,
C          SLOPE B,STANDARD ERROR IN REGRESSION LINE SE, AND XB
C          AND SX ARE STORED IN THE IROW ROW OF AN ARRAY
C          COR(I,J) TO BE USED LATER IN THE SUBROUTINES WHICH
C          DRAW SAMPLES FROM THE DEPENDENT VARIABLES.
      IF(LT.NE.0) GO TO 3
      FN = FLOAT(MM)
      S1 = 0.0
      S2 = 0.0
      DO 1 I=1,MM
      S1 = S1+XSR(I)
      S2 = S2+YSR(I)
    1 CONTINUE
      XB = S1/FN
      YB = S2/FN
      S1 = 0.0
      S2 = 0.0
      S = 0.0
      DO 2 I=1,MM
      S1 = S1+(XSR(I)-XB)**2
      S2 = S2+(YSR(I)-YB)**2
      S = S+(XSR(I)-XB)*(YSR(I)-YB)
    2 CONTINUE
      VX = S1/(FN-1.0)
      VY = S2/(FN-1.0)
      SX = SQRT(VX)
      SY = SQRT(VY)
      B = S/S1
      A = YB-XB*B
      R = (B*SX)/SY
      GO TO 4
    3 READ(1,10) R,XB,SX,YB,SY
   10 FORMAT(8E10.3)
      B=(R*SY)/SX
      A=YB-XB*B
    4 R2=R**2
      ARG=1.-R2
      SE=SY*SQRT(ARG)
      COR(IROW,1)=A
      COR(IROW,2)=B
      COR(IROW,3)=SE
      COR(IROW,4)=XB
      COR(IROW,5)=SX
      RETURN
      END
```

- -

```
      SUBROUTINE WSAMP(NGROUP,IX,ICOL,C1,C2)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51)
C          THIS SUBROUTINE DRAWS A SAMPLE OF SIZE NGROUP FROM
C          THE WEIBULL DISTRIBUTION FUNCTION.  THE PARAMETERS
C          A AND B TO BE USED FOR THE WEIBULL ARE OBTAINED
C          FROM THE ARRAY PAR(I,J).  THE DISTRIBUTION HAS
C          LOWER AND UPPER CLIPS C1 AND C2 RESPECTIVELY
C          (SUPPLIED FROM THE MAINLINE).  THE SAMPLE IS
C          STORED IN THE ICOL COLUMN OF THE ARRAY VAR(I,J).
C          THE METHOD USED IN SAMPLING IS THE DIRECT APPROACH.
      A=PAR(ICOL,1)
      B=PAR(ICOL,2)
      P=1.0/B
      DO 1 I=1,NGROUP
    2 CALL RANDU(IX,IY,R)
      IX=IY
      IF(R.EQ.1.0) GO TO 2
      R=(C2-C1)*R+C1
      ARG=1.0/(1.0-R)
      VAR(I,ICOL)=(ALOG(ARG)/A)**P
    1 CONTINUE
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE NORSAM(NGROUP,IX,ICOL,C1,C2)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51)
C         THIS SUBROUTINE DOES PRECISELY THE SAME THING AS
C         THE WSAMP SUBROUTINE EXCEPT THAT IT DRAWS THE
C         GIVEN SAMPLE FROM THE NORMAL DISTRIBUTION INSTEAD
C         OF THE WEIBULL.  THE DIRECT APPROACH IS ALSO USED
C         HERE.
      XBAR=PAR(ICOL,1)
      SD=PAR(ICOL,2)
      TPI=2.0*3.14159
      DO 1 I=1,NGROUP
    2 CALL RANDU(IX,IY,R1)
      IX=IY
      IF(R1.EQ.0.0) GO TO 2
      ARG=1.0/(R1*R1)
      ARG=ALOG(ARG)
      TX=SQRT(ARG)
      CALL RANDU(IX,IY,R2)
      IX=IY
      R2=R2*TPI
      VAR(I,ICOL)=XBAR+SD*TX*SIN(R2)
      IF((VAR(I,ICOL).LT.C1).OR.(VAR(I,ICOL).GT.C2)) GO TO 2
      IF(ICODE(ICOL).NE.2) GO TO 1
      IF(VAR(I,ICOL).LT.0.0) GO TO 2
    1 CONTINUE
      RETURN
      END
```

- -

```
      SUBROUTINE LOGSAM(NGROUP,IX,ICOL,C1,C2)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51)
C         THIS SUBROUTINE SAMPLES FROM THE LOG-NORMAL
C         DISTRIBUTION IN PRECISELY THE SAME MANNER AS
C         NORSAM AND WSAMP DO FROM THE NORMAL AND WEIBULL
C         RESPECTIVELY.  IT ALSO USES THE DIRECT APPROACH.
      CALL NORSAM(NGROUP,IX,ICOL,C1,C2)
      DO 1 I=1,NGROUP
      VAR(I,ICOL)=EXP(VAR(I,ICOL)
    1 CONTINUE
      RETURN
      END
```

- -

```
      SUBROUTINE GAMSAM(NGROUP,IX,ICOL)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51)
C         THIS SUBROUTINE SAMPLES FROM THE GAMMA DISTRIBUTION
C         IN THE SAME MANNER AS THE ABOVE SUBROUTINES SAMPLE
C         FROM THEIR RESPECTIVE DISTRIBUTIONS.  THE ONLY
C         DIFFERENCE IS THAT THIS SUBROUTINE USES THE
C         REJECTION METHOD FOR SAMPLING FROM THE GAMMA.
      ALPHA=PAR(ICOL,1)
      BETA=PAR(ICOL,2)
      XU=BETA*(ALPHA+5.0*SQRT(ALPHA))
      AM1=ALPHA-1.0
      CALL GAMMA(ALPHA,GA,I)
      CON=GA*(BETA**ALPHA)
      YU=((AM1)**AM1)/(GA*BETA*EXP(AM1))
      DO 1 I=1,NGROUP
    2 CALL RANDU(IX,IY,XVAL)
      IX=IY
      CALL RANDU(IX,IY,YVAL)
      IX=IY
      XVAL=XVAL*XU
      YVAL=YVAL*YU
      FXVAL=GDF(XVAL,ALPHA,BETA,CON)
      IF(YVAL.LE.FXVAL) GO TO 3
      GO TO 2
    3 VAR(I,ICOL)=XVAL
    1 CONTINUE
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      FUNCTION GDF(X,ALPHA,BETA,CON)
C          THIS SUBPROGRAM EVALUATES THE PROBABILITY DENSITY
C          FUNCTION FOR THE GAMMA DISTRIBUTION AT THE POINT
C          X.  THE GAMMA PARAMETERS ARE ALPHA AND BETA AND
C          THE CONSTANT CON IS GIVEN BY CON=GA*(BETA**ALPHA)
C          WHERE GA IS THE GAMMA FUNCTION EVALUATED AT ALPHA.
      ARG1=-(X/BETA)
      ARG2=ALPHA-1.0
      GDF=((X**ARG2)*EXP(ARG1))/CON
      RETURN
      END
```

- -

```
      SUBROUTINE GAMMA(XX,GX,IER)
C          THIS SUBROUTINE EVALUATES THE GAMMA FUNCTION AT
C          THE POINT XX.  THE VALUE IS STORED AND RETURNED IN
C          THE LOCATION GX.
      IF(XX-34.5)6,6,4                                                  GAMMAM01
    4 IER=2                                                             GAMMAM02
      GX = 1.0E18
      RETURN                                                            GAMMAM04
    6 X=XX                                                              GAMMAM05
      ERR=1.0E-6
      IER=0
      GX=1.0
      IF(X-2.0)50,50,15
   10 IF(X-2.0)110,110,15
   15 X=X-1.0
      GX=GX*X
      GO TO 10                                                          GAMMA 10
   50 IF(X-1.0)60,120,110
C         SEE IF X IS NEAR NEGATIVE INTEGER OR ZERO                     GAMMA 12
   60 IF(X-ERR)62,62,80
   62 K=X
      Y=FLOAT(K)-X
      IF(ABS(Y)-ERR)130,130,64
   64 IF(1.0-Y-ERR)130,130,70
C         X NOT NEAR A NEGATIVE INTEGER OR ZERO                         GAMMA 18
   70 IF(X-1.0)80,80,110
   80 GX=GX/X
      X=X+1.0
      GO TO 70                                                          GAMMA 22
  110 Y=X-1.0
      GY=1.0+Y*(-0.5771017+Y*(+0.9858540+Y*(-0.8764218+Y*(+0.8328212+
     1Y*(-0.5684729+Y*(+0.2548205+Y*(-0.05149930)))))))
      GX=GX*GY
  120 RETURN                                                            GAMMA 27
  130 IER=1
      RETURN                                                            GAMMA 29
      END                                                               GAMMA 30
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE LCSAMP(NGROUP,NPASS,IX,ICOL,IVAR,IDVAR)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51),IDCODE(10,2),
     1COR(10,5)
C          THIS SUBROUTINE DRAWS A SAMPLE OF SIZE NGROUP FOR
C          A DEPENDENT INPUT VARIABLE FROM THE LOG-NORMAL
C          DISTRIBUTION.  IVAR IS THE SUBSCRIPT OF THE
C          INDEPENDENT VARIABLE WITH WHICH THE GIVEN DEPENDENT
C          VARIABLE IS CORRELATED.  IDVAR IS THE NUMBER OF
C          THE DEPENDENT VARIABLE.  ICOL IS THE COLUMN OF THE
C          STORAGE ARRAY VAR(I,J) IN WHICH THE SAMPLE IS
C          STORED.  THE METHOD USED IS THE ONE DISCUSSED IN
C          THE DESCRIPTION OF THIS PROGRAM.
      YINT=COR(IDVAR,1)
      SLOPE=COR(IDVAR,2)
      SE=COR(IDVAR,3)
      XBARS=COR(IDVAR,4)
      SDS=COR(IDVAR,5)
      FN=NGROUP*NPASS
      TPI=2.0*3.14159
      DO 30 I=1,NGROUP
      AVAR=ALOG(VAR(I,IVAR))
      XBAR=YINT+(SLOPE*AVAR)
      T1=((AVAR-XBARS)*(AVAR-XBARS))/(FN*SDS*SDS)
      SD=SE*(1.0+1.0/FN+T1)
   31 CALL RANDU(IX,IY,R1)
      IX=IY
      IF(R1.EQ.0.0) GO TO 31
      ARG=1.0/(R1*R1)
      ARG=ALOG(ARG)
      TX=SQRT(ARG)
      CALL RANDU(IX,IY,R2)
      IX=IY
      R2=R2*TPI
      VAR(I,ICOL)=XBAR+(SD*TX*SIN(R2))
      VAR(I,ICOL)=EXP(VAR(I,ICOL))
   30 CONTINUE
      RETURN
      END
```

- -

```
      SUBROUTINE NCSAMP(NGROUP,NPASS,IX,ICOL,IVAR,IDVAR)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51),IDCODE(10,2),
     1COR(10,5)
C          THIS SUBROUTINE DOES THE SAME THING AS LCSAMP
C          ABOVE EXCEPT THAT IT DRAWS THE SAMPLE FOR THE
C          DEPENDENT VARIABLE FROM A NORMAL DISTRIBUTION.
      YINT=COR(IDVAR,1)
      SLOPE=COR(IDVAR,2)
      SE=COR(IDVAR,3)
      XBARS=COR(IDVAR,4)
      SDS=COR(IDVAR,5)
      FN=NGROUP*NPASS
      TPI=2.0*3.14159
      DO 30 I=1,NGROUP
      XBAR=YINT+(SLOPE*VAR(I,IVAR))
      T1=((VAR(I,IVAR)-XBARS)*(VAR(I,IVAR)-XBARS))/(FN*SDS*SDS)
      SD=SE*(1.0+1./FN+T1)
   31 CALL RANDU(IX,IY,R1)
      IX=IY
      IF(R1.EQ.0.0) GO TO 31
      ARG=1.0/(R1*R1)
      ARG=ALOG(ARG)
      TX=SQRT(ARG)
      CALL RANDU(IX,IY,R2)
      IX=IY
      R2=R2*TPI
      VAR(I,ICOL)=XBAR+(SD*TX*SIN(R2))
      IF(VAR(I,ICOL).LT.0.0) GO TO 31
   30 CONTINUE
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE RANDU(IX,IY,YFL)
C          THIS SUBROUTINE GENERATES A PSEUDO-RANDOM NUMBER
C          BETWEEN 0 AND 1 AND RETURNS IT IN THE PARAMETER
C          YFL.  IT ALSO RETURNS A VALUE OF IY WHICH IS TO BE
C          USED AS IX IN THE NEXT PASS THROUGH THIS SUBROUTINE.
      IY=IX*899
      IF(IY) 5,6,6
    5 IY=IY+32767+1
    6 YFL=IY
      YFL=YFL/32767.
      RETURN
      END
```

```
      SUBROUTINE       BACKT(NGROUP,ICOL1,CY1,CY2,TRAN,XI1,XI2,ANGL)
C          THIS SUBROUTINE PERFORMS THE
C          TRANSFORMATION BACK TO ORIGINAL
C          DATA VALUES AFTER LINEAR
C          TRANSFORMATION PER PARA 19.8 OF
C          STATISTICAL THEORY WITH ENGINEERING
C          APPLICATIONS, HALD, WILEY, 1967.
C          THE TRANSFORMATION MAKES
C          CORRELATION COEFFICIENT ZERO OF
C          DATA PAIRS OF PARTIALLY CORRELATED
C          DATA WITH MEANS OF ZERO.  A SECOND
C          TRANSFORMATION IS REQUIRED TO
C          ADD A CONSTANT TO THE DATA TO
C          MAKE IT POSITIVE SO THAT LOGS
C          CAN BE TAKEN.
C
C          WRITTEN BY LEE MOTE - 3/10/76
C
C          XI1 IS MEAN OF X
C          XI2 IS MEAN OF Y
C          ANGL IS ANGLE OF ROTATION IN
C          RADIANS
C          Y1 IS TRANSFORMED ARRAY FOR
C          INDEPENDENT VARIABLE
C          Y2 IS TRANSFORMED ARRAY FOR
C          DEPENDENT VARIABLE
C          X1 IS RESTORED VALUES OF
C          INDEPENDENT VARIABLE
C          X2 IS RESTORED VALUES OF
C          DEPENDENT VARIABLE
      BIT TRAN(16)
      DIMENSION X1(51), X2(51),Y1(51),Y2(51)
      COMMON ICODE(10),VAR(51,10),PAR(10,2),SEV(51),IDCODE(10,2),
     1COR(10,5),XSR(250),YSR(250)
      N=NGROUP
      ICOL2=ICOL1+1
      DO 1 I=1,N
      Y1(I)=VAR(I,ICOL1)-CY1
      Y2(I)=VAR(I,ICOL2)-CY2
      IF(TRAN(2))Y1(I)=ALOG(VAR(I,ICOL1))-CY1
      IF(TRAN(3))Y2(I)=ALOG(VAR(I,ICOL2))-CY2
1     CONTINUE
      DO 2 I=1,N
      X1(I)=XI1+Y1(I)*COS(ANGL)-Y2(I)*SIN(ANGL)
      X2(I)=XI2+Y1(I)*SIN(ANGL)+Y2(I)*COS(ANGL)
      IF(TRAN(2))X1(I)=EXP(X1(I))
      IF(TRAN(3)) X2(I)=EXP(X2(I))
2     CONTINUE
C
      DO 3 I=1,N
      VAR(I,ICOL1)=X1(I)
3     VAR(I,ICOL2)=X2(I)
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      FUNCTION SF(IVAL)
      COMMON ICODE(10).VAR(51.10).PAR(10.2).SEV(51)
C         THIS SUBPROGRAM EVALUATES THE OUTPUT VARIABLE SF
C         AT A PARTICULAR VALUE IVAL ASSOCIATED WITH THE
C         INPUT VARIABLE ARRAY VAR(I.J).
      SF=(30.1243*VAR(IVAL.1)*VAR(IVAL.2))/(VAR(IVAL.3)*VAR(IVAL.3))
      RETURN
      END
```

```
      SUBROUTINE SEVCAL(NGROUP.XMIN.XMAX.SMEAN.SSQ)
      COMMON ICODE(10).VAR(51.10).PAR(10.2).SEV(51)
C         THIS SUBROUTINE CALCULATES THE NGROUP VALUES OF
C         SEVERITY (OR OUTPUT VARIABLE) AND STORES THEM IN
C         THE ARRAY SEV(I) FOR LATER USE.  IT ALSO KEEPS
C         TRACK OF A RUNNING SUM OF SEVERITIES. SUM OF THE
C         SQUARES OF SEVERITIES. AND AN OVERALL XMIN AND
C         XMAX VALUE FOR SEVERITY.
      DO 1 I=1.NGROUP
      SEV(I)=SF(I)
      SMEAN=SMEAN+SEV(I)
      SSQ=SSQ+(SEV(I)**2)
      IF(SEV(I).LT.XMIN) XMIN=SEV(I
      IF(SEV(I).GT.XMAX) XMAX=SEV(I)
    1 CONTINUE
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE CINT(NCINT.XMIN.XMAX.NINT.XE.W.NCFLAG)
      DIMENSION XE(35)
C          THIS SUBROUTINE ESTABLISHES THE CLASS INTERVALS TO
C          BE USED IN THE PROGRAM.  THIS SUBROUTINE IS
C          BRANCHED TO AFTER THE FIRST PASS OF NGROUP VALUES
C          OF THE OUTPUT VARIABLE SEV ARE CALCULATED.  IT
C          THEN USES THESE VALUES AND STURGES' RULE FOR
C          SETTING UP THE CLASS INTERVALS OR IT READS THE
C          NECESSARY DATA AND SETS UP THE CLASS INTERVALS
C          THAT THE USER DESIRES TO HAVE.
      IF(NCFLAG.NE.0) GO TO 2
      XN=FLOAT(NCINT)
      NINT=1.5+3.3*ALOG10(XN)
      IF (XMAX.GT.50.0) XMAX = 50.0
      W=(XMAX-XMIN)/FLOAT(NINT)
      W=W+.0001*W
      GO TO 3
    2 READ(1.5)NINT.XMIN.W
    5 FORMAT(I5.2F10.3)
    3 XE(1) = XMIN
      K=NINT+1
      DO 1 I=2.K
      XE(I)=XE(I-1)+W
    1 CONTINUE
      RETURN
      END
```

- -

```
      SUBROUTINE AFREQ(NGROUP.NINT.XE.AF.EMP)
      DIMENSION XE(35).AF(35).EMP(3)
      COMMON ICODE(10).VAR(51.10).PAR(10.2).SEV(51)
C          THIS SUBROUTINE ACCEPTS NGROUP VALUES OF SEVERITY
C          SEV (OR OUTPUT VARIABLE) AND SEPARATES THEM INTO
C          THE NINT CLASS INTERVALS WITH ENDPOINTS IN THE
C          ARRAY XE.  THE ACTUAL FREQUENCIES ARE COUNTED IN
C          THE REAL ARRAY AF.  FURTHER. THE NUMBER OF
C          OBSERVED VALUES OF SEV BELOW .1. BETWEEN .1 AND 1.
C          AND ABOVE 1 ARE COUNTED AND STORED IN THE ARRAY
C          EMP(I). I=1.2.3.
      DO 1 J=1.NGROUP
      IF(SEV(J).GE.XE(1)) GO TO 2
      AF(1)=AF(1)+1.0
      GO TO 1
    2 DO 3 I=1.NINT
      IF((SEV(J).GE.XE(I)).AND.(SEV(J).LT.XE(I+1))) GO TO 4
      GO TO 3
    4 AF(I+1)=AF(I+1)+1.0
      GO TO 1
    3 CONTINUE
      AF(NINT+2)=AF(NINT+2)+1.0
    1 CONTINUE
      DO 5 J=1.NGROUP
      IF(SEV(J).LE..1) GO TO 7
      IF(SEV(J).GT.1.0) GO TO 8
      EMP(2)=EMP(2)+1.0
      GO TO 5
    7 EMP(1)=EMP(1)+1.0
      GO TO 5
    8 EMP(3)=EMP(3)+1.0
    5 CONTINUE
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE OUTPUT(SMEAN.SDEV.PROB.SNUM.PNUM.AF.CF.XSAMP.XPOP.
     1NINT.XE.ITIL.XMIN.XMAX)
      DIMENSION PROB(3).SNUM(3).PNUM(3).AF(35).CF(35).XE(35).ITIL(20)
C          THIS SUBROUTINE PRINTS THE OUTPUT OF THE PROGRAM
C          INCLUDING MEAN. STANDARD DEVIATION. ETC..
      WRITE(5.10)(ITIL(I).I=1.20)
   10 FORMAT(1H1.40X.20A2)
      WRITE(5.1) SMEAN.SDEV
      WRITE(5.2) PROB(1).PROB(2).PROB(3)
      WRITE(5.3) XSAMP.SNUM(1).SNUM(2).SNUM(3)
      WRITE(5.4) XPOP.PNUM(1).PNUM(2).PNUM(3)
      WRITE(5.5)
      K=NINT+1
      XLE=XMIN
      WRITE(5.7) XLE.XE(1).AF(1).CF(1)
      DO 6 I = 2.K
    6 WRITE(5.7) XE(I-1).XE(I).AF(I).CF(I)
      XRE=XMAX
      WRITE(5.7) XE(NINT+1).XRE.AF(NINT+2).CF(NINT+2)
    1 FORMAT(1H0.'THE SAMPLE MEAN IS '.E14.7//
     1' THE SAMPLE STANDARD DEVIATION IS '.E14.7)
    2 FORMAT(1H0.' PROB(S.LE..1) = '.E14.7/
     1' PROB(.1.LT.S.LE.1.0) = '.E14.7/' PROB(S.GT.1.0) = '.E14.7)
    3 FORMAT(1H0///' XSAMP = '.E14.7/' SNUM(1) = '.E14.7/
     1' SNUM(2) = '.E14.7/' SNUM(3) = '.E14.7)
    4 FORMAT(1H0///' XPOP = '.E14.7/' PNUM(1) = '.E14.7/
     1' PNUM(2) = '.E14.7/' PNUM(3) = '.E14.7)
    5 FORMAT(1H0//////3X.'CLASS INTERVALS'.19X.'ACTUAL FREQUENCY'
     120X.'CUMULATIVE FREQUENCY')
    7 FORMAT(1H0.'FROM '.E11.4.2X.'TO '.E11.4.10X.E14.7.20X.E14.7)
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE C1PLOT(N,XBAR,SD,NINT,W,XE,XMIN,XMAX,CFREQ,ITIL,LT)
C         THIS SUBROUTINE DRAWS A CUMULATIVE FREQUENCY PLOT
C         FOR A GIVEN SET OF DATA.  THE VARIABLES AND THEIR
C         DESCRIPTION ARE GIVEN BELOW:
C     N     =NUMBER OF OBSERVATIONS
C     NINT  =NUMBER OF CLASS INTERVALS
C     XMIN  =MINIMUM VALUE
C     XMAX  =MAXIMUM VALUE
C     W     =WIDTH OF CLASS INTERVALS
C     XE    =VECTOR OF ENDPOINTS OF CLASS INTERVALS
C            BEGINNING WITH XE(1)=XMIN AND GOING TO
C            XE(NINT+1)=XMAX
C     KOUNT =VECTOR CONTAINING THE NUMBER OF OBSERVATIONS
C            THAT FALL IN THE VARIOUS CLASS INTERVALS
C     CFREQ =VECTOR GIVING THE CUMULATIVE FREQUENCIES
C     ITIL  =CAPTION OR HEADING BELOW X-AXIS
C     LT    =LENGTH OF THE X-AXIS TITLE IN ACTUAL LETTERS
C     XBAR  =MEAN OF SAMPLE
C     SD    =STANDARD DEVIATION OF SAMPLE
      DIMENSION XE(35),CFREQ(35),ITIL(20)
      IPLOT=6
      I=MSCFA(IPLOT,'PL')
      FN=FLOAT(N)
      NV=NINT+1
      DO 1 I=1,NV
    1 CFREQ(I)=CFREQ(I)*FN
      CALL PLOTS(IDUM,IDUM,IPLOT)
      CALL PLOT(1.0,0.0,-2)
      XINT=FLOAT(NINT)+1.0
      CALL AXIS(0.0,0.0,ITIL,-LT,XINT,0.0,XE(1),W)
      FACT=.9
      CALL FACTOR(FACT)
      DELTAV=FLOAT(N/10)+1.
      CALL AXIS(0.,0.,'NUMBER OF PLANTS',16,10.,90.,0.,DELTAV)
      CALL FACTOR(1.0)
      CALL PLOT(0.0,10.0*FACT,3)
      CALL PLOT(XINT,10.0*FACT,2)
      XE(NINT+2)=XE(1)
      XE(NINT+3)=W
      CFREQ(NINT+2)=0.0
      CFREQ(NINT+3)=DELTAV/FACT
      CALL LINE(XE,CFREQ,NINT+1,1,1,11)
      CALL SYMBOL(XINT,4.0,.21,'SAMPLE SIZE = ',0.0,14)
      XN=FLOAT(N)
      CALL NUMBER(999.0,999.0,.21,XN,0.0,-1)
      CALL SYMBOL(XINT,3.5,.21,'MIN. VALUE = ',0.0,13)
      CALL NUMBER(999.0,999.0,.21,XMIN,0.0,2)
      CALL SYMBOL(XINT,3.0,.21,'MAX. VALUE = ',0.0,13)
      CALL NUMBER(999.0,999.0,.21,XMAX,0.0,2)
      CALL SYMBOL(XINT,2.5,.21,'MEAN = ',0.0,7)
      CALL NUMBER(999.0,999.0,.21,XBAR,0.0,2)
      CALL SYMBOL(XINT,2.0,.21,'STD. DEV. = ',0.0,12)
      CALL NUMBER(999.0,999.0,.21,SD,0.0,2)
      CALL PLOT(0.0,0.0,-3)
      CALL PLOT(30.0,0.0,-3)
      RETURN
      END
```

Table C-3 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS

```
      SUBROUTINE FDPLOT(N,XBAR,SD,NINT,W,X,XMIN,XMAX,KOUNT,ITIL,LT)
C         THIS SUBROUTINE DRAWS A FREQUENCY HISTOGRAM FOR A
C         GIVEN SET OF DATA.  THE VARIABLES AND THEIR
C         DESCRIPTION ARE GIVEN BELOW:
C     N      =NUMBER OF OBSERVATIONS
C     NINT   =NUMBER OF CLASS INTERVALS
C     XMIN   =MINIMUM VALUE
C     XMAX   =MAXIMUM VALUE
C     W      =WIDTH OF CLASS INTERVALS
C     XE     =VECTOR OF ENDPOINTS OF CLASS INTERVALS
C              BEGINNING WITH XE(1)=XMIN AND GOING TO
C              XE(NINT+1)=XMAX
C     KOUNT  =VECTOR CONTAINING THE NUMBER OF OBSERVATIONS
C              THAT FALL IN THE VARIOUS CLASS INTERVALS
C     CFREQ  =VECTOR GIVING THE CUMULATIVE FREQUENCIES
C     ITIL   =CAPTION OR HEADING BELOW X-AXIS
C     LT     =LENGTH OF THE X-AXIS TITLE IN ACTUAL LETTERS
C     XBAR   =MEAN OF SAMPLE
C     SD     =STANDARD DEVIATION OF SAMPLE
      DIMENSION XE(35),KOUNT(35),ITIL(20),YK(35),X(35)
      IPLOT=6
      I=MSCFA(IPLOT,'PL')
      NINT=NINT-1
      DO 3 I=1,20
    3 XE(I)=X(I)
      CALL PLOTS(IDUM,IDUM,IPLOT)
      CALL PLOT(1.0,0.0,-2)
      XINT=FLOAT(NINT)+1.0
      CALL AXIS(0.0,0.0,ITIL,-LT,XINT,0.0,XE(1),W)
      KK=NINT+2
      DO 1 K=1,KK
      M=KOUNT(K)
    1 YK(K)=FLOAT(M)
      YK(NINT+3)=0.0
      CALL SCALE(YK,8.0,NINT+3,1)
      YK(NINT+4)=YK(NINT+5)
      CALL AXIS(-1.0,0.0,'NUMBER OF OBSERVATIONS',22,8.0,90.0,
     1YK(NINT+3),YK(NINT+4))
      YVAL=YK(1)/YK(NINT+4)
      CALL SYMBOL(-.5,YVAL,.08,11,0.0,-1)
      LDL=NINT+1
      DO 2 I=1,LDL
      XE(I)=XE(I)+W/2.0
    2 YK(I)=YK(I+1)
      XE(NINT+2)=XE(1)-W/2.0
      XE(NINT+3)=W
      YK(NINT+2)=0.0
      YK(NINT+[illegible])=YK(NINT+4)
      YVAL=YK(1)/YK(NINT+3)
      CALL PLOT(.5,YVAL,2)
      CALL LINE(XE,YK,NINT+1,1,1,11)
      CALL SYMBOL(XINT,4.0,.21,'SAMPLE SIZE = ',0.0,14)
      XN=FLOAT(N)
      CALL NUMBER(999.0,999.0,.21,XN,0.0,-1)
      CALL SYMBOL(XINT,3.5,.21,'MIN. VALUE = ',0.0,13)
      CALL NUMBER(999.0,999.0,.21,XMIN,0.0,2)
      CALL SYMBOL(XINT,3.0,.21,'MAX. VALUE = ',0.0,13)
      CALL NUMBER(999.0,999.0,.21,XMAX,0.0,2)
      CALL SYMBOL(XINT,2.5,.21,'MEAN = ',0.0,7)
      CALL NUMBER(999.0,999.0,.21,XBAR,0.0,2)
```

- -

```
      CALL SYMBOL(XINT,[illegible],.21,'STD. DEV. = ',0.0,12)
      CALL NUMBER(999.0,999.0,.21,SD,0.0,2)
      CALL PLOT(0.0,0.0,-3)
      CALL PLOT(50.0,0.0,-4)
      NINT=NINT+1
      RETURN
      END
```

APPENDIX D

THE GOODNESS-OF-FIT PROGRAM[5,7-12]

The goodness-of-fit program is designed to take sample data from some population with an unknown distribution and "fit" the data to various continuous distributions by one of the standard procedures of statistics. The program will print out various sample parameters, such as the mean, standard deviation, coefficient of skewness, and measure of kurtosis. It will also print out the corresponding parameters for the theoretical continuous distributions which are fitted to the data. Finally, using Sturge's rule to set up class intervals, the program will calculate the chi-square value to be used in a chi-square test for goodness-of-fit as discussed earlier. In addition, using the right endpoints of the class intervals as comparison points, the program calculates the residual sum of the squares of the difference in the theoretical cumulative distribution and the actual cumulative distribution. All of the calculations indicated above can be used to determine how well the given theoretical distribution fits the actual data.

---

[7] Mendenhall, W., and R. L. Scheaffer. Mathematical Statistics with Applications. North Scituate, Duxbury Press, 1973.

[8] Walpole, R. E., and R. H. Myers. Probability and Statistics for Engineers and Scientists. New York, The MacMillan Co., 1972.

[9] Siegel, S. Nonparametric Statistics. New York, McGraw-Hill Book Co., 1956.

[10] Duncan, A. J. Quality Control and Industrial Statistics. Chicago, Richard D. Irwin, Inc., 1952.

[11] Cramer, H. Mathematical Methods of Statistics. Princeton, Princeton University Press, 1946. 575 p.

[12] Olt, W. R., and D. T. Magee. Random Sampling as an Inexpensive Means for Measuring Average Annual Air Pollutant Concentrations in Urban Areas. (Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston. June 15-20, 1975.

1. THEORETICAL DISTRIBUTIONS USED AND THE METHODOLOGY UTILIZED FOR FITTING THEM TO THE DATA

There are four distributions which are used in the goodness-of-fit program: the normal distribution, the Weibull distribution, the gamma distribution, and the log-normal distribution. These represent a wide range of continuous distributions and are, therefore, deemed sufficient to cover most (if not all) of the data encountered.

Three methods are routinely used in statistics for fitting data to a continuous distribution: the method of maximum likelihood, the method of moments, and the method of least squares. The theoretical aspects of these methods are discussed in the statistics texts listed in the references at the end of this report. The applications of these methods to the specific distributions under consideration are discussed below.

First, the data are fitted to the Weibull distribution by two methods: the method of maximum likelihood and the method of least squares. For the method of maximum likelihood, the following system of equations (in the parameters a and b) need to be solved:

$$\left\{ \begin{array}{l} \frac{n}{a} - \sum_{i=1}^{n} X_i^{\,b} = 0 \\ \\ \frac{n}{b} + \sum_{i=1}^{n} \log X_i - a \sum_{i=1}^{n} X_i^{\,b} \log X_i = 0 \end{array} \right\} \begin{array}{r} \text{(D-1)} \\ \\ \text{(D-2)} \end{array}$$

where n = sample size and the $\{X_i\}_{i=1}^{n}$ are the sample data values. Eliminating "a" from the system we obtain the following equation in b alone:

$$\frac{\sum_{i=1}^{n} X_i^{\,b} \log X_i^{\,b}}{\sum_{i=1}^{n} X_i^{\,b}} - \frac{1}{n}\sum_{i=1}^{n} \log X_i^{\,b} - 1 = 0 \qquad \text{(D-3)}$$

The program uses a numerical scheme called the method of false position to solve the above equation for b. Then it obtains "a" by a substitution of b into Equation D-1. Thus, the necessary parameters, a and b, in the Weibull distribution function are obtained.

For the method of least squares, the right endpoints of the class intervals are taken as the x-values and their cumulative distribution values are taken as the y-values to be transformed by the usual double-log transformation and used in the method of least squares for obtaining "a" and b. Since the cumulative distribution value for the right endpoint of the last class interval is 1.0 and this value cannot be allowed in the double-log transformation, log (log $\frac{1}{1-Y}$), the Y-value for the last class interval is set at 0.999.

The next fit considered is the normal distribution. The sample mean, $\overline{X}$, and the (unbiased) sample standard deviation are taken as the values for μ and σ in the normal distribution function. This is "almost" a maximum likelihood fit to the normal. Actually, the maximum likelihood estimators of μ and σ are the sample mean and the biased sample standard deviation. However, the unbiased standard deviation should work as well or better.

The next distribution used is the gamma distribution. The data are fitted to this distribution by the method of moments (since this method is easy to apply to the gamma). The method of moments yields the system of equations below (in the parameters α and β) to be solved:

$$\alpha\beta = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(D-4)}$$

$$\alpha\beta^2 = \frac{\Sigma(X_i-\overline{X})^2}{n} \quad \text{(biased variance used in method of moments)} \qquad \text{(D-5)}$$

Squaring the first equation and dividing it by the second one obtains:

$$\alpha = \frac{\left(\frac{\Sigma X_i}{n}\right)^2}{\frac{\Sigma(X_i-\overline{X})^2}{n}} = \frac{n\ \overline{X}^2}{\Sigma(X_i-\overline{X})^2} \qquad \text{(D-6)}$$

The program uses the above equation to obtain α directly and subsequently substitute this value into the first equation to obtain $\beta = \frac{\overline{X}}{\alpha}$ directly.

Finally, consider fitting the log-normal distribution to the data. If the data are to be log-normally distributed, the natural logarithms of the data points should be approximately normal. Hence, to perform a log-normal fit to the data, the logarithm of each data point is obtained and used as data for which to perform a maximum likelihood fit to the normal as described earlier. That is, the mean of the sample of log values is used as μ and the unbiased sample standard deviation of the log values is used as σ for the normal distribution.

The above remarks conclude the discussion of the theoretical distributions used and the methods for fitting the data to each of them individually.

2. FORM AND DESCRIPTION OF OUTPUT

The first thing printed out is a title describing the data which are to be analyzed. The sample statistics are then printed out, including $\overline{X}$, standard deviation, m_3 (third central moment), g_1 (coefficient of skewness), etc. The various distributions are subsequently fitted one by one in the manner described above and the information described in the introduction to this section is printed out for each of the fits. The first fit is the Weibull Maximum Likelihood Fit and the last one is the Log-Normal Fit. In the section for each fit, the class intervals and the corresponding theoretical frequency and actual frequency of the data in these class intervals are printed out. In calculating the value of chi-square, the program automatically pools frequency classes on the upper and lower tails until they obey "the rule of 5" for the chi-square test. In calculating the number of degrees of freedom, the program automatically reduces the number of class intervals to take into account the pooling of the class intervals described above.

One should keep in mind, however, that the program does nothing about a class interval with a theoretical frequency less than five unless it occurs as the first or last class interval. In that case, the program simply pools it with the class interval below or above it until "the rule of 5" is obeyed. Since most distributions (including the theoretical ones being discussed) will generally not have small frequency class intervals in their center (except for bimodal ones), this procedure will for the most part take care of any necessary reduction to obey "the rule of 5." However, if more reduction is deemed necessary, this can easily be done by hand, since the class intervals and the frequencies will be printed out in the output. As an example of the above, consider the following table of class intervals and frequencies:

Table D-1. THEORETICAL AND ACTUAL FREQUENCIES FOR VARIOUS CLASS INTERVALS

Class interval	Theoretical frequency	Actual frequency
1	82.1	84
2	53.8	57
3	35.6	34
4	22.2	14
5	13.2	14
6	7.7	6
7	4.3	7
8	2.4	5
9	2.8	3

The program will pool the last two class intervals producing eight class intervals and, therefore, 8-3 = 5 degrees of freedom. Class interval 7 does not obey "the rule of 5" either. However, since it is close to five, the error produced in leaving it alone and adding one more degree of freedom to the test is small. Furthermore, if it were pooled with class interval 6 above it, the resulting value of chi-square would be less but so also would the number of degrees of freedom and the final conclusion would be nearly the same.

3. FORM OF INPUT

The input to the program is divided into two parts. The first part consists of a single card containing three pieces of information about the data to be analyzed, namely:

N = sample size (in cols 1-5, integer and right justified)

ITIL(I),I=1,30 = title of the data (in cols 6-65 anywhere)

NFLAG = variable telling whether this is the last data set to be analyzed by the program on this run. If NFLAG = 0 (i.e., the card is blank in col 80), then the program expects to analyze another data set after completing this one. If NFLAG = 1 (i.e., the card has a 1 in col 80), this signals the last data set and the program will terminate after analyzing the current data.

The second part of the input contains the values of the data punched eight per card under an F10.3 format until the data are exhausted. The format for reading the data into the program could easily be changed, if desired.

A listing of the program and output pertaining to the coal-fired electric utility data is given in Appendix C for reference purposes.

Table D-2. COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      DIMENSION XE(20),AF(20),XSR(20),YSR(20),XES(20),AFS(20),RF(20)
      DIMENSION ITIL(30)
      COMMON DATA(500),XMIN,XMAX,XBAR,SDEV
C          THIS PROGRAM IS DESIGNED TO TAKE SAMPLE DATA FROM
C          SOME POPULATION WITH AN UNKNOWN DISTRIBUTION AND
C          "FIT" THIS DATA TO VARIOUS CONTINUOUS DISTRIBUTIONS
C          BY ONE OF THE STANDARD PROCEDURES OF STATISTICS.
C          THE PROGRAM WILL PRINT OUT VARIOUS SAMPLE PARAMETERS
C          SUCH AS MEAN, STANDARD DEVIATION,COEFFICIENT OF
C          SKEWNESS, AND MEASURE OF KURTOSIS.  IT WILL ALSO
C          PRINT OUT THE CORRESPONDING PARAMETERS FOR THE
C          THEORETICAL CONTINUOUS DISTRIBUTIONS WHICH ARE
C          FITTED TO THE DATA.  FINALLY, USING STURGE'S RULE
C          TO SET UP CLASS INTERVALS, THE PROGRAM WILL
C          CALCULATE THE CHI-SQUARE VALUE TO BE USED IN A
C          CHI-SQUARE TEST FOR GOODNESS-OF-FIT.  IN ADDITION,
C          USING THE RIGHT ENDPOINTS OF THE CLASS INTERVALS
C          AS COMPARISON POINTS, THE PROGRAM CALCULATES THE
C          RESIDUAL SUM OF THE SQUARES OF THE DIFFERENCE IN
C          THE THEORETICAL CUMULATIVE DISTRIBUTION AND THE
C          ACTUAL CUMULATIVE DISTRIBUTION.
    5 READ(1,1) N,(ITIL(I),I = 1,30),NFLAG
    1 FORMAT(I5,30A2,10X,I5)
      READ(1,2)(DATA(I),I = 1,N)
    2 FORMAT(8F10.3)
      WRITE(5,10)(ITIL(I),I=1,30)
   10 FORMAT(1H1,30X,30A2)
      CALL SSTAT(N)
      CALL CINT(N,XMIN,XMAX,NINT,W,XES,AFS)
C   WEIBULL MAXIMUM LIKELIHOOD FIT
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL SOLVE(N,A,B)
      IF(B.LT.0.0)GO TO 3
      CALL TSKEW(A,B)
      CALL RFWEIB(NINT,XE,A,B,RF)
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL CHITST(N,NINT,RF,AF,XE)
C    WEIBULL LEAST SQUARES FIT
    3 CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL SETUP(N,NINT,XE,AF,XSR,YSR)
      CALL SR(NINT,XSR,YSR,A,B)
      CALL TSKEW(A,B)
      CALL RFWEIB(NINT,XE,A,B,RF)
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL CHITST(N,NINT,RF,AF,XE)
C    NORMAL FIT
      WRITE(5,20)
   20 FORMAT(1H1/////53X,'NORMAL FIT')
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL RFNORM(NINT,XE,XBAR,SDEV,RF)
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL CHITST(N,NINT,RF,AF,XE)
C          GAMMA METHOD OF MOMENTS FIT
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL GSOLVE(XBAR,SDEV,ALPHA,BETA,N)
      IF(ALPHA.LT.1.0) GO TO 50
      CALL RFGAM(NINT,XE,ALPHA,BETA,RF)
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL CHITST(N,NINT,RF,AF,XE)
      GO TO 51
- - - - - - - - - - - - - - - - - - - - - - - - - -
   50 WRITE(5,52)
   52 FORMAT(1H0,'THE GAMMA DISTRIBUTION WILL NOT FIT THIS DATA')
C    LOG NORMAL FIT
   51 WRITE(5,100)
  100 FORMAT(1H1/////53X,'LOG NORMAL FIT')
      DO 101 I=1,N
      DATA(I) = ALOG(DATA(I))
  101 CONTINUE
      CALL SSTAT(N)
      CALL CINT(N,XMIN,XMAX,NINT,W,XES,AFS)
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL RFNORM(NINT,XE,XBAR,SDEV,RF)
      CALL RELOAD(XES,AFS,XE,AF,NINT)
      CALL CHITST(N,NINT,RF,AF,XE)
      IF(NFLAG.EQ.0) GO TO 5
      CONTINUE
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE RFNORM(NINT.XE.XBAR.SDEV.RF)
      DIMENSION XE(20).RF(20)
C          THIS SUBROUTINE IS DESIGNED TO CALCULATE THE
C          THEORETICAL RELATIVE FREQUENCY RF OF THE NORMAL
C          DISTRIBUTION WITH MEAN XBAR AND STANDARD DEVIATION
C          SDEV OVER THE CLASS INTERVALS WHOSE ENDPOINTS ARE
C          GIVEN IN THE ARRAY XE.  NINT IS THE NUMBER OF
C          CLASS INTERVALS REPRESENTED.
      DO 1 I=1.NINT
    1 RF(I)=0.0
      IF((XBAR-3.*SDEV).LT.XE(2)) GO TO 2
      XE(1) = 0.0
      GO TO 3
    2 XE(1) = XBAR-3.*SDEV
    3 M = NINT-1
      S = 0.0
      DO 6 I =1.M
      DX = (XE(I+1)-XE(I))/1000.0
      CALL INTEG(XE(I).XE(I+1).DX.XBAR.SDEV.RF(I))
      S = S+RF(I)
    6 CONTINUE
      RF(NINT) = 1.0-S
      RETURN
      END
```

- -

```
      SUBROUTINE INTEG(C.D.DX.XBAR.SDEV.V)
C          THIS SUBROUTINE CALCULATES THE INTEGRAL OF THE
C          NORMAL PROBABILITY DENSITY FUNCTION PDF WITH MEAN
C          XBAR AND STANDARD DEVIATION SDEV OVER THE INTERVAL
C          C TO D IN INCREMENTS OF DX BY THE TRAPEZOIDAL
C          RULE.  1000 ITERATIONS ARE USED.
      V = 0.0
      A2 = SQRT(2.*3.14159)*SDEV
      Y1 = PDF(C.XBAR.SDEV.A2)
      Y2 = Y1
      DO 1 I = 1.1000
      XI = FLOAT(I)
      X2 = C + XI*DX
      Y1 = Y2
      Y2 = PDF(X2.XBAR.SDEV.A2)
    1 V= V+ .5*(Y1+Y2)*DX
      RETURN
      END
```

- -

```
      FUNCTION PDF(X.XBAR.SDEV.A2)
C          THIS FUNCTION EVALUATES THE NORMAL DENSITY
C          FUNCTION WITH MEAN XBAR AND STANDARD DEVIATION
C          SDEV AT THE POINT X.  THE ARGUMENT A2 IS A
C          CONSTANT TRANSFERRED IN WITH VALUE --
C          A2=SQRT(2*PI)*SDEV.
      A1 = ((X-XBAR)/SDEV)**2
      ARG = -.5*A1
      PDF = EXP(ARG)/A2
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE SETUP(N,NINT,XE,AF,XSR,YSR)
      DIMENSION XE(20),AF(20),XSR(20),YSR(20)
C         THIS SUBROUTINE ACCEPTS THE DATA AFTER IT IS
C         ARRANGED INTO CLASS INTERVALS AND ACTUAL
C         FREQUENCIES HAVE BEEN CALCULATED AND SETS UP THE
C         RIGHT ENDPOINTS OF THE CLASS INTERVALS WITH THEIR
C         RESPECTIVE CUMULATIVE FREQUENCIES FOR THE LEAST
C         SQUARES FIT TO THE WEIBULL.
      S = 0.0
      XN = FLOAT(N)
      DO 1 I=1,NINT
      XSR(I) = XE(I+1)
      S = S+AF(I)
    1 YSR(I) = S/XN
      YSR(NINT)=.999
      DO 2 I=1,NINT
      XSR(I) = ALOG(XSR(I))
      ARG = 1.0/(1.0-YSR(I))
      ARG = ALOG(ARG)
      YSR(I) = ALOG(ARG)
    2 CONTINUE
      RETURN
      END
```

- -

```
      FUNCTION F(N,B)
C         THIS FUNCTION SUBPROGRAM CALCULATES THE EQUATIO
C         VALUE AT A POINT B FOR THE EQUATION USED IN
C         SOLVING FOR B IN THE WEIBULL MAXIMUM LIKELIHOOD
C         FIT.  N IS THE TOTAL NUMBER OF DATA POINTS.
      DIMENSION XB(500),ALXB(500)
      COMMON X(500)
      S1=0.0
      S2=0.0
      S3=0.0
      DO 1 I=1,N
      XB(I)=X(I)**B
      ALXB(I)=ALOG(XB(I))
      S1=S1+XF(I)*ALXB(I)
      S2=S2+XB(I)
    1 S3=S3+ALXB(I)
      FN=FLOAT(N)
      FN=1/FN
      F=(S1/S2)-FN*S3-1.0
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE TSKEW(A.B)
C          THIS SUBROUTINE CALCULATES AND PRINTS THE
C          THEORETICAL VALUES OF THE MEAN. VARIANCE. STANDAR
C          DEVIATION. THIRD AND FOURTH CENTRAL MOMENTS.
C          COEFFICIENT OF SKEWNESS. AND MEASURE OF KURTOSIS
C          FOR THE WEIBULL DISTRIBUTION FUNCTION WITH
C          PARAMETERS A AND B GIVEN.  IT ALSO EVALUATES AND
C          PRINTS THE I AND J POINTS OF THE FUNCTION FOR I=1.
C          ---. 5 AND J=99. ---. 95.
      Y1=1.0+(1.0/B)
      Y2=1.0+(2.0/B)
      Y3=1.0+(3.0/B)
      Y4=1.0+(4.0/B)
      CALL GAMMA(Y1.G1.I)
      CALL GAMMA(Y2.G2.J)
      CALL GAMMA(Y3.G3.K)
      CALL GAMMA(Y4.G4.L)
      A1 = A**(-1.0/B)
      A2=A**(-2.0/B)
      A3=A**(-3.0/B)
      A4=A**(-4.0/B)
      XMU = A1*G1
      SIGMA2=(G2-G1**2)*A2
      SIGMA=SQRT(SIGMA2)
      X3=(G3-3.*G1*G2+2.0*G1**3)*A3
      X4=(G4-4.*G1*G3+6.0*G2*G1**2.0-3.*G1**4.0)*A4
      SIGMA3=SIGMA**3.0
      SIGMA4=SIGMA**4.0
      GAMMA1=X3/SIGMA3
      GAMMA2=X4/SIGMA4
      WRITE(5.1) GAMMA1.GAMMA2
    1 FORMAT(1H /'COEFFICIENT OF SKEWNESS = '.F10.3.10X.'MEASURE OF KURT
     1OSIS = '.F10.3)
      WRITE(5.2)XMU.SIGMA2.SIGMA.X3.X4
    2 FORMAT(' MEAN = '.E14.7.5X.'VARIANCE = '.E14.7.5X.'STAN. DEV. = '.
     1E14.7//' THIRD CENTRAL MOMENT = '.E14.7.5X.
     1'FOURTH CENTRAL MOMENT = '.E14.7)
      P=1.0/B
      DO 3 I=1.5
      K=100-I
      XP1=FLOAT(I)/100.0
      XP2=1.0-XP1
      ARG1=1.0/(1.0-XP1)
      ARG2=1.0/(1.0-XP2)
      P1=(ALOG(ARG1)/A)**P
      P2=(ALOG(ARG2)/A)**P
      WRITE(5.4)I.K.P1.P2
    4 FORMAT(1H0.'THE '.I2.' AND '.I2.' POINTS ARE '.E14.7.' AND '.
     1E14.7)
    3 CONTINUE
      RETURN
      END
```

```
      FUNCTION WF(X.A.B)
C          THIS FUNCT ON SUBPROGRAM EVALUATES THE CUMULATIVE
C          WEIBULL DISTRIBUTION FUNCTION WITH PARAMETERS A
C          AND B AT THE POINT X.
      ARG=-(A*(X**B))
      WF=1.0-EXP(ARG)
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE RFWEIB(NINT,XE,A,B,RF)
      DIMENSION XE(20),RF(20),CF(20)
C          THIS SUBROUTINE CALCULATES THE RELATIVE FREQUENCY
C          RF OVER THE CLASS INTERVALS WHOSE ENDPOINTS ARE
C          THE ARRAY XE FOR THE WEIBULL DISTRIBUTION FUNCTION
C          WHOSE PARAMETERS ARE A AND B.  NINT IS THE NUMBER
C          OF CLASS INTERVALS.
      MM = NINT-1
      DO 5 I = 1,MM
      XVAL = XE(I+1)
      CF(I) = WF(XVAL,A,B)
    5 CONTINUE
      CF(NINT) = 1.0
      RF(1) = CF(1)
      DO 6 I = 2,NINT
    6 RF(I) = CF(I)-CF(I-1)
      RETURN
      END
```

- -

```
      SUBROUTINE SR(NINT,XSR,YSR,A,B)
C          THIS SUBROUTINE CALCULATES THE PARAMETERS A AND B
C          FOR THE WEIBULL DISTRIBUTION FUNCTION BY THE LEAST
C          SQUARES METHOD.  THE ALREADY SETUP ARRAYS XSR AND
C          YSR ARE SUPPLIED TO THE SUBROUTINE.  NINT IS THE
C          NUMBER OF CLASS INTERVALS.
      DIMENSION XSR(20),YSR(20)
      FN=FLOAT(NINT)
      S1 = 0.0
      S2 = 0.0
      DO 1 I=1,NINT
      S1 = S1+XSR(I)
      S2 = S2+YSR(I)
    1 CONTINUE
      XB = S1/FN
      YB = S2/FN
      S1 = 0.0
      S2 = 0.0
      S = 0.0
      DO 2 I=1,NINT
      S1 = S1+(XSR(I)-XB)**2
      S2 = S2+(YSR(I)-YB)**2
      S = S+(XSR(I)-XB)*(YSR(I)-YB)
    2 CONTINUE
      VX = S1/(FN-1.0)
      VY = S2/(FN-1.0)
      SX = SQRT(VX)
      SY = SQRT(VY)
      B = S/S1
      AL = YB-XB*B
      A = EXP(AL)
      R = (B*SX)/SY
      R2 = R**2
      ARG = (S2*(1.-R2))/(FN-2.0)
      SE = SQRT(ARG)
      SB = SE/SQRT(S1)
      TV = B/SB
      WRITE(5,400)
  400 FORMAT(1H0,////53X,'WEIBULL LEAST SQUARES FIT')
      WRITE(5,10)A,B
   10 FORMAT(1H0,'WEIBULL PARAMETERS ARE G = 0.0,   A = ',E14.4,
     15X,'B = ',E14.4)
      WRITE(5,11)R,R2,SE,SB,TV
   11 FORMAT(1H0,'R = ',E12.4,5X,'RSQ = ',E12.4,5X,'SE = ',E12.4,
     15X,'SB = ',E12.4,5X,'TV = ',E12.4)
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE RELOAD(XES,AFS,XE,AF,NINT)
      DIMENSION XES(20),AFS(20),XE(20),AF(20)
C         THE ARRAYS XE OF CLASS INTERVAL ENDPOINTS AND AF
C         OF ACTUAL FREQUENCIES ARE ALTERED IN CERTAIN PARTS
C         OF THIS PROGRAM TO FIT THE NEEDS AT THE TIME.
C         THUS, RATHER THAN RECALCULATE THEM EVERYTIME THEY
C         ARE NEEDED, THEY WERE STORED IN SAVED ARRAYS XES
C         AND AFS UPON INITIAL CALCULATION.  THIS SUBROUTINE
C         SIMPLY RELOADS THE WORKING ARRAYS XE AND AF WITH
C         THEIR SAVED VALUES.
      DO 1 I = 1,NINT
      XE(I) = XES(I)
      AF(I) = AFS(I)
    1 CONTINUE
      XE(NINT+1) = XES(NINT+1)
      RETURN
      END
```

- -

```
      SUBROUTINE GSOLVE(XBAR,SDEV,ALPHA,BETA,N)
          THIS SUBROUTINE FITS THE GAMMA DENSITY FUNCTION TO
C         THE GIVEN DATA BY THE METHOD OF MOMENTS.  XBAR AND
C         SDEV ARE THE MEAN AND STANDARD DEVIATION OF THE
C         GIVEN DATA WHILE ALPHA AND BETA ARE THE CALCULATED
C         PARAMETERS FOR THE GAMMA DENSITY FUNCTION.  N IS
C         THE TOTAL NUMBER OF DATA POINTS IN OUR SAMPLE.  IN
C         ADDITION, THIS SUBROUTINE CALCULATES AND PRINTS
C         THE THEORETICAL VALUES OF THE MEAN, STANDARD
C         DEVIATION, COEFFICIENT OF SKEWNESS, ETC. FOR THE
C         FITTED GAMMA FUNCTION.
C
      FN=FLOAT(N)
      ALPHA=(FN*XBAR*XBAR)/((FN-1.0)*SDEV*SDEV)
      BETA=XBAR/ALPHA
      X3=2.0*ALPHA*BETA*BETA*BETA
      G1=2.0/SQRT(ALPHA)
      X4=((3.0*ALPHA**2)+(6.0*ALPHA))*BETA**4
      G2=3.0+6./ALPHA
      VAR=SDEV*SDEV
      WRITE(5,1)
    1 FORMAT(1H1/////53X,'GAMMA METHOD OF MOMENTS FIT')
      WRITE(5,2) ALPHA,BETA
    2 FORMAT(1H0,'GAMMA PARAMETERS ARE:  ','ALPHA = ',E14.7,5X,'BETA = '
     1,E14.7)
      WRITE(5,3)XBAR,VAR,SDEV,X3,X4,G1,G2
    3 FORMAT(1H0,'XBAR = ',E14.7,5X,'VAR = ',E14.7,5X,'SDEV = ',
     1E14.7,5X//'THIRD CENTRAL MOMENT = ',E14.7,5X,'FOURTH CENTRAL MOMEN
     2T = ',E14.7,5X//'COEFFICIENT OF SKEWNESS = ',E14.7,5X,
     3'MEASURE OF KURTOSIS = ',E14.7)
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE RFGAM(NINT,XE,ALPHA,BETA,RF)
      DIMENSION XE(20),RF(20)
C          THIS SUBROUTINE SETS UP THE ARRAY OF RELATIVE
C          FREQUENCIES RF FOR THE GAMMA DENSITY FUNCTION WITH
C          PARAMETERS ALPHA AND BETA.  XE IS THE ARRAY OF
C          CLASS INTERVAL ENDPOINTS AND NINT IS THE NUMBER OF
C          CLASS INTERVALS.
      DO 1 I=1,NINT
    1 RF(I)=0.0
      XE(1)=1.0E-05
      M = NINT-1
      S = 0.0
      DO 4 I = 1,M
      DX=(XE(I+1)-XE(I))/1000.0
      CALL GINTEG(XE(I),XE(I+1),DX,ALPHA,BETA,RF(I))
      S = S+RF(I)
    4 CONTINUE
      RF(NINT) = 1.0 - S
      RETURN
      END
```

```
      SUBROUTINE GINTEG(C,D,DX,ALPHA,BETA,V)
C          THIS SUBROUTINE INTEGRATES THE GAMMA DENSITY
C          FUNCTION WITH PARAMETERS ALPHA AND BETA OVER THE
C          INTERVAL C TO D.  IT USES 1000 ITERATIONS OF THE
C          TRAPEZOIDAL RULE WHERE THE INCREMENT SIZE IS DX.
C          THE VALUE V OF THE INTEGRAL IS RETURNED TO RFGAM
C          AS THE RELATIVE FREQUENCY OF THE GAMMA DENSITY FOR
C          THE INTERVAL C TO D
      V=0.0
      CALL GAMMA(ALPHA,GA,1)
      CON=(BETA**ALPHA)*GA
      Y1=GDF(C,ALPHA,BETA,CON)
      Y2=Y1
      DO 1 I=1,1000
      XI=FLOAT(I)
      X2=C+XI*DX
      Y1=Y2
      Y2=GDF(X2,ALPHA,BETA,CON)
      V=V+.5*(Y2+Y1)*DX
    1 CONTINUE
      RETURN
      END
```

```
      FUNCTION GDF(X,ALPHA,BETA,CON)
C          THIS SUBPROGRAM EVALUATES THE GAMMA DENSITY
C          FUNCTION WITH PARAMETERS ALPHA AND BETA AT THE
C          POINT X.  CON IS A CONSTANT SUPPLIED TO GDF HAVING
C          VALUE (BETA**ALPHA)*GA WHERE GA IS THE GAMMA
C          FUNCTION EVALUATED AT ALPHA.
      ARG1=-(X/BETA)
      ARG2=ALPHA-1.0
      GDF=((X**ARG2)*EXP(ARG1))/CON
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE SSTAT (N)
      COMMON X(500), XMIN, XMAX, XBAR, SDEV
C          THIS SUBROUTINE CALCULATES AND PRINTS THE SAMPLE
C          STATISTICS FOR THE GIVEN DATA.
      S=0.0
      DO 1 I=1,N
    1 S=S+X(I)
      XBAR=S/FLOAT(N)
      S2=0.0
      S3=0.0
      S4=0.0
      DO 2 I=1,N
      S2=S2+(X(I)-XBAR)**2
      S3=S3+(X(I)-XBAR)**3
    2 S4=S4+(X(I)-XBAR)**4
      VAR=S2/FLOAT(N-1)
      SDEV=SQRT(VAR)
      FN=FLOAT(N)
      XM3=(FN/((FN-1.0)*(FN-2.0)))*S3
      XT1=((FN**2.0-2.*FN+3.0)/((FN-1.0)*(FN-2.0)*(FN-3.0)))*S4
      XT2=((3.0*(FN-1)*(2.*FN-3.0))/(FN*(FN-2.0)*(FN-3.0)))*VAR**2
      XM4=XT1-XT2
      G1=XM3/SDEV**3
      G2=XM4/SDEV**4
      XMIN=X(1)
      XMAX=X(1)
      DO 3 I=1,N
      IF(X(I).LT.XMIN) XMIN=X(I)
      IF(X(I).GT.XMAX) XMAX=X(I)
    3 CONTINUE
      WRITE(5,10)
   10 FORMAT(52X,'SAMPLE STATISTICS')
      WRITE(5,11) XBAR,VAR,SDEV,XM3,XM4,G1,G2,XMIN,XMAX
   11 FORMAT(1H0,'XBAR = ',E14.7,5X,'VAR =',E14.7,5X,'SD = ',E14.7,5X//
     1' M3 = ',E14.7,5X,'M4 = ',E14.7,//' G1 = ',E14.7,5X,'G2 = ',
     1E14.7,5X,'XMIN = ',E14.7,5X,'XMAX = ',E14.7)
      RETURN
      END
```

```
      SUBROUTINE CINT(N,XMIN,XMAX,NINT,W,XE,AF)
      DIMENSION XE(20),AF(20)
      COMMON X(500)
C          THIS SUBROUTINE SETS UP THE CLASS INTERVALS FOR
C          THE GIVEN DATA USING STURGE'S RULE TO DETERMINE
C          THE NUMBER - NINT - OF CLASS INTERVALS.  IT ALSO
C          SORTS THE ORIGINAL DATA INTO THESE CLASS INTERVALS
C          COUNTING THE NUMBER IN EACH BY USE OF THE ARRAY AF
C          OF ACTUAL FREQUENCIES.
      XN = FLOAT(N)
      NINT = 1.5 + 3.3 *ALOG10(XN)
      W = (XMAX - XMIN)/FLOAT(NINT)
      W = W + .00001*W
      XE(1) = XMIN
      K = NINT + 1
      DO 10 I = 2,K
   10 XE(I) = XE(I-1) + W
      DO 1 I=1,NINT
    1 AF(I)=0.0
      DO 2 J=1,N
      DO 3 I = 1,NINT
      IF(X(J).GE.XE(I).AND.X(J).LT.XE(I+1)) GO TO 4
      GO TO 3
    4 AF(I)=AF(I)+1.0
      GO TO 2
    3 CONTINUE
    2 CONTINUE
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE CHITST(N,NINT,RF,AF,XE)
      DIMENSION AF(20),RF(20),TF(20),XE(20)
C          THIS SUBROUTINE CALCULATES THE THEORETICAL
C          FREQUENCIES FOR THE CLASS INTERVALS AND PRINTS THE
C          TABLE OF CLASS INTERVALS WITH THEIR RESPECTIVE
C          THEORETICAL AND ACTUAL FREQUENCIES.  IT ALSO
C          CALCULATES AND PRINTS THE VALUE OF SSE, THE SUM OF
C          THE SQUARES OF THE ERROR.  FINALLY, IT CALCULATES
C          AND PRINTS THE VALUE OF CHI-SQUARE AND ITS
C          ASSOCIATED DEGREES OF FREEDOM FOR THIS DATA.
      FN = FLOAT(N)
      DO 7 I=1,NINT
    7 TF(I)=RF(I)*FN
      WRITE(5,10)
   10 FORMAT(1H0,'   CLASS INTERVALS',20X,'THE ORETICAL FREQUENCY',20X,
     1'ACTUAL FREQUENCY')
      DO 11 I=1,NINT
   11 WRITE(5,12) XE(I),XE(I+1),TF(I),AF(I)
   12 FORMAT(1H0,'FROM ',F10.2,' TO ',F10.2,15X,F10.3,30X,F10.3)
      S1 = 0.0
      S2 = 0.0
      SSE = 0.0
      DO 30 I = 1,NINT
      S1 = S1 + RF(I)
      S2 = S2 +(AF(I)/FN)
      SSE = SSE + (S1-S2)**2
   30 CONTINUE
      WRITE(5,31) SSE
   31 FORMAT(1H0,'SSE = ',F16.7)
      K=NINT
      L=1
      DO 20 I=1,NINT
      IF(TF(I).GE.5.0) GO TO 21
      TF(I+1)=TF(I+1)+TF(I)
      AF(I+1)=AF(I+1)+AF(I)
      L=L+1
   20 CONTINUE
   21 DO 22 I=1,NINT
      IF(TF(NINT-I+1).GE.5.0) GO TO 23
      TF(NINT-I)=TF(NINT-I)+TF(NINT-I+1)
      AF(NINT-I)=AF(NINT-I)+AF(NINT-I+1)
      K=K-1
   22 CONTINUE
   23 S=0.0
      DO 24 I=L,K
   24 S=S+((TF(I)-AF(I))**2.0)/TF(I)
      NOF=K-L-2
      WRITE(5,50) S,NOF
   50 FORMAT(1H0//'CHI-SQUARE = ',F15.6,' WITH ',I3,' DEGREES OF FREEDOM
     1')
      RETURN
      END
```

Table D-2 (continued). COMPUTER LISTINGS OF THE SIMULATION PROGRAMS - GOODNESS-OF-FIT PROGRAM

```
      SUBROUTINE SOLVE(N,A,B)
      COMMON XX(500)
C           THIS SUBROUTINE SOLVES  OR THE PARAMETERS A AND B
C           IN THE WEIBULL DISTRIBUTION FUNCTON BY THE METHOD
C           OF MAXIMUM LIKELIHOOD.  IT USES THE METHOD OF
C           FALSE POSITION IN THE SOLUTION PROCEDURE.  N IS
C           THE TOTAL NUMBER OF DATA POINTS AND XX IS THE
C           ARRAY OF DATA POINTS.
      X1 = 1.0
      X2 = 2.0
      Y1 = F(N,X1)
      Y2 = F(N,X2)
   71 IF((Y1.LT.0.0.AND.Y2.GT.0.0).OR.(Y1.GT.0.0.AND.Y2.LT.0.0)) GO TO
     170
      X1 = X2
      Y1 = Y2
      X2 = X2 + 1.0
      Y2 = F(N,X2)
      IF(X1.GT.10.0) GO TO 2
      GO TO 71
    2 WRITE(5,11)
   11 FORMAT(' BETA IS > 10.0 OR < 1.0 OR THERE IS AN EVEN NUMBER OF
     1ROOTS BETWEEN TWO INTEGER VALUES')
      B=-1.0
      RETURN
   70 IF(Y1.LT.0.0.AND.Y2.GT.0.0) GO TO 75
      K = 1
      L = 2
      GO TO 3
   75 K = 1
      L = 1
    3 X=X1-Y1*((X2-X1)/(Y2-Y1))
      Y = F(N,X)
      IF(ABS(Y).LE..00001) GO TO 100
      IF((Y.LT.0.0.AND.L.EQ.1).OR.(Y.GT.0.0.AND.L.EQ.2)) GO TO 5
      X2 = X
      Y2 = Y
   20 K = K + 1
      IF(K.GE.5000) GO TO 101
      GO TO 3
    5 X1 = X
      Y1 = Y
      GO TO 20
  101 WRITE(5,12)
   12 FORMAT(' THE NUMBER OF ITERATIONS EXCEEDED 5000')
  100 B=X
      S=0.0
      DO 50 I=1,N
   50 S=S+XX(I)**B
      A=FLOAT(N)/S
      WRITE(5,400)
  400 FORMAT(1H0/////50X,'WEIBULL MAXIMUM LIKELIHOOD FIT')
      WRITE(5,500) A,B
  500 FORMAT(1H0,'WEIBULL PARAMETERSS ARE G = 0.0,     A = ',E14.4,
     1'B = ',E14.4)
      RETURN
      END
```

APPENDIX E

TREATMENT OF CORRELATED DATA BY LINEAR TRANSFORMATION

1. THEORETICAL DISCUSSION OF LINEAR TRANSFORMATION

Dependent (correlated) input variables are discussed in Section IV.B.2. If the original variables are binormally distributed, by linear transformation of variables, it is possible to find two normally distributed and stochastically independent linear functions of these variables.[13] Assume that X1 and X2 are correlated variables. With the introduction of variables Y1 and Y2, we can transform the (X1, X2) variables to the (Y1, Y2) coordinate system which will have its origin in the point (X1, X2) = (ξ1, ξ2) and the Y-axis will form angle α with the X-axis.

$$Y1 = (X1 - \xi 1) \cos \alpha + (X2 - \xi 2) \sin \alpha \qquad \text{(E-1)}$$

$$Y2 = -(X1 - \xi 1) \sin \alpha + (X2 - \xi 2) \cos \alpha \qquad \text{(E-2)}$$

The correlation coefficient of (Y1, Y2) depends on the angle α. The following equation is used to find a value for α for which the correlation coefficient is zero.

$$\tan 2\alpha = \frac{2\rho S_1 S_2}{S_1{}^2 - S_2{}^2} \text{ for } S_1 \neq S_2 \text{ and } \alpha = \frac{\pi}{4} \text{ for } S_1 = S_2 \qquad \text{(E-3)}$$

where ρ = correlation coefficient of (X1, X2) data pairs
S_1 = sample standard deviation of X1 data
S_2 = sample standard deviation of X2 data

The data values (Y1, Y2) may be restored to the original values (X1, X2) by the following inverse transformations:

---

[13] Hald, A. Statistical Theory with Engineering Applications. New York, John Wiley & Sons, Inc., 1952. p. 596-599.

$$X1 = \xi 1 + Y1 \cos \alpha - Y2 \sin \alpha \qquad \text{(E-4)}$$

$$X2 = -\xi 2 + Y1 \sin \alpha + Y2 \cos \alpha \qquad \text{(E-5)}$$

The above equations were implemented using APL language on a time-sharing terminal. Using the population data for coal consumption and stack heights, the (X1, X2) data pairs were correlated with $\rho = 0.559$. After transformation the correlation coefficient $\rho = (2.978E - 17)$ which is essentially zero.

2. SIMULATION IMPLEMENTATION

The variables coal consumed (CC) and stack height (H) were treated as log-normal distributions. The transformation of the 24 sample values for CC and H were performed on the APL terminal as previously discussed. The correlation coefficient for CC with H was 0.3705. After transformation, the correlation coefficient was 9.9E-17, which is essentially zero. Table E-1 lists the inputs and results of various simulation runs.

Equations E-4 and E-5 were implemented in a subroutine in the simulation program to restore the transformed data to actual values (see Table C-3 for a listing of the program), with the result that the mean severity was 9.31. The simulation was repeated with stack height considered as an independent variable. The mean value of severity was 11.25. The t-test was made to see if these values were significantly different from the mean severity of 9.25 arrived at earlier in this report.[14]

---

[14] Alder, H. L., and Roessler, E. B. Introduction to Probability and Statistics. San Francisco, W. H. Freeman and Company, 1968. p. 136-140.

Table E-1. COMPARISON OF SIMULATION RESULTS

Method	Parameter	Mean	Standard deviation	Comment
Section IV.B.2	Coal consumed[a]	1326	1329	A = 0.9487E - 3 B = 0.9856
	Stack heights[b]	93.6	36	Dependent variable
	Percent sulfur[c]	1.82	1.15	A = 0.3039 B = 1.667
	Resultant severity[a]	9.25	12.48	
Variables treated as independent	Stack heights[d]	4.472	0.3751	Log-normal distribution
	Resultant severity[a]	11.25	19.17	
Transformation of variables	Coal consumed[e]	0.0	0.9731	Transformed
	Stack heights[e]	0.0	0.3442	Transformed
	Resultant severity	9.31	12.0	

[a] Coal consumption in kg/yr; stack height in m; severity is dimensionless.

[b] Stack height treated as a dependent variable correlated with coal consumed.

[c] Same value used for each simulation.

[d] Stack height treated as an independent variable.

[e] Coal consumed and stack heights linearly transformed to make correlation coefficient zero; both were treated as log-normal distributions.

The t-test assumes that the variate X is normally distributed with mean σ. If all possible samples of n variates are taken from this population and their means are denoted by $\overline{X}$, then

$$t = \frac{\overline{X}-\mu}{S_{\overline{X}}} \quad \text{(E-6)}$$

$$S = \hat{S}\sqrt{\frac{n}{n-1}} \quad \text{(E-7)}$$

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} \quad \text{(E-8)}$$

where t = value from table of the Student t-distribution with n-1 degrees of freedom

$\overline{X}$ = mean of X

μ = population mean

$S_{\overline{X}}$ = standard deviation of $\overline{X}$

$\hat{S}$ = sample standard deviation

S = standard deviation corrected for small sample size

n = sample size

In the sample calculation as shown below, it is assumed that the population mean of 9.25 is from the simulation results of Section IV.B.2. From a reference table, t(23) = 2.807 at the 99% confidence level.

$$S_{\overline{X}} = \frac{19.17}{\sqrt{23}} \left(\sqrt{\frac{24}{23}}\right) = 3.998$$

$$t = \frac{11.25 - 9.25}{3.998} = 0.50$$

Since the above t value is less than 2.807, it is concluded that the means are from the same population. This is also true for the mean of the simulation using the linear transformation.

The F-distribution allows one to test whether the variances of two means are from the same population.[15] A confidence level such as 99% (significance level of 1%) is selected. A null hypothesis of $(\sigma_1)^2 = (\sigma_2)^2$ is made. The alternative is $(\sigma_1)^2 \neq (\sigma_2)^2$ and the samples are not from the same populations with the same variance. Usually, the larger variance is divided by the smaller:

---

[15] Koosis, D. J. Statistics. New York, John Wiley and Sons, Inc., 1972. p. 155-160.

$$F = \frac{(S_1)^2}{(S_2)^2} \text{ where } S_1 > S_2$$

For the case of treating stack height as an independent variable,

$$F = \frac{(19.17)^2}{(12.48)^2} = \frac{367}{156} = 2.35$$

and for the linear transformation case,

$$F = \frac{(12.0)^2}{(12.48)^2} = \frac{144}{156} = 0.923$$

For a sample size of 24, the critical region is F >2.72 (found from a table). Thus, it is concluded that the variances are not significantly different at the 1% confidence level. It thus appears that linear transformation is a valid method to use with correlated data.

SECTION VII

REFERENCES

1. Slade, D. H. (ed.). Meteorology and Atomic Energy. Environmental Science Services Administration, Air Resources Labs. Silver Spring. AEC Publication No. TID-24190. July 1968. 445 p.

2. Turner, D. B. Workbook of Atmospheric Dispersion Estimates, 1970 Revision. U.S. Department of Health, Education, and Welfare. Cincinnati. Public Health Service Publication No. 999-AP-26. May 1970. 84 p.

3. Parzen, E. Modern Probability Theory and Its Applications. New York, John Wiley & Sons, 1960.

4. Springer, C. H., et al. Probabilistic Models. Homewood, Richard D. Irwin, Inc., 1968.

5. Curran, T. C., and N. H. Frank. Assessing the Validity of the Lognormal Model when Predicting Maximum Air Pollution Concentrations. (Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston. June 15-20, 1975.)

6. Geary, R. C., and E. S. Pearson. Tests of Normality. London, University College, 1938.

7. Mendenhall, W., and R. L. Scheaffer. Mathematical Statistics with Applications. North Scituate, Duxbury Press, 1973.

8. Walpole, R. E., and R. H. Myers. Probability and Statistics for Engineers and Scientists. New York, The MacMillan Co., 1972.

9. Siegel, S. Nonparametric Statistics. New York, McGraw-Hill Book Co., 1956.

10. Duncan, A. J. Quality Control and Industrial Statistics. Chicago, Richard D. Irwin, Inc., 1952.

Preceding page blank

11. Cramer, H. Mathematical Methods of Statistics. Princeton, Princeton University Press, 1946. 575 p.

12. Olt, W. R., and D. T. Magee. Random Sampling as an Inexpensive Means for Measuring Average Annual Air Pollutant Concentrations in Urban Areas. (Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston. June 15-20, 1975.)

13. Hald, A. Statistical Theory with Engineering Applications. New York, John Wiley & Sons, Inc., 1952. p. 596-599.

14. Alder, H. L., and Roessler, E. B. Introduction to Probability and Statistics. San Francisco, W. H. Freeman and Company, 1968. p. 136-140.

15. Koosis, D. J. Statistics. New York, John Wiley and Sons, Inc., 1972. p. 155-160.